Ana Filipa Antunes Viriato Silva

Licenciatura em Matemática

Modelação do Risco de Crédito numa Carteira de Crédito ao Consumo

Dissertação para obtenção do Grau de Mestre em Matemática e Aplicações, no ramo de Actuariado, Estatística e Investigação Operacional

Orientadores : Professora Doutora Gracinda Rita Diogo Guerreiro, Professora Auxiliar, Faculdade de Ciências e Tecnologias, UNL, Portugal
Professor Doutor Manuel Leote Tavares Inglês Esquível, Professor Associado, Faculdade de Ciências e Tecnologias, UNL, Portugal

Júri:

Presidente: Professor Doutor Jorge Orestes Lasbarrères Cerdeira

Arguente: Professor Doutor José Faias

Vogal: Professor Doutor Manuel Leote Tavares Inglês Esquível

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro, 2014

Modelação do Risco de Crédito numa Carteira de Crédito ao Consumo

Modelação do Risco de Crédito numa Carteira de Crédito ao Consumo

Dissertação para obtenção do Grau de Mestre em
Matemática e Aplicações, no ramo de Actuariado, Estatística e
Investigação Operacional

Ana Filipa Antunes Viriato Silva

Licenciatura em Matemática

Agradecimentos

Agradeço aos meus orientadores, Professora Doutora Gracinda Rita Guerreiro e Professor Doutor Manuel L. Esquível por todos os conhecimentos valiosos que me transmitiram, pelo apoio, pela experiência, pela dedicação e por toda a disponibilidade ao longo desta dissertação. E pelo convite na participação na construção de um artigo, foi muito gratificante.

Ao Professor Doutor José Fernandes, pela sua disponibilidade no tratamento de dados e à instituição bancária que nos forneceu os dados para que fosse possível realizar esta dissertação.

Aos meus colegas e amigos deste percurso pela amizade, pelo apoio, pelo incentivo constante e por todos os momentos que foram passados em conjunto.

Por último e com máxima importância, à minha família pelo carinho, pela paciência e pelo apoio incondicional, que sem ele nunca teria chegado ao fim deste grande projecto. E por estarem sempre por perto nas alturas em que era preciso um gigante empurrão.

Resumo

A análise de risco de crédito nas instituições bancárias e a mensuração do risco é de extrema importância para as instituições, uma vez que a concessão de crédito é a sua principal actividade. A capacidade de distinguir "bom" e "mau" cliente é um processo decisivo na constituição do crédito, pelo que são aplicados modelos de *Credit Scoring* , modelos quantitativos que consistem numa análise estatística à qualidade do crédito.

O objectivo desta dissertação é estimar a probabilidade de incumprimento de cada cliente em função das variáveis sócio-económicas e demográficas, tendo por base dados de uma carteira de crédito ao consumo de uma Instituição Bancária de Cabo Verde, através de uma técnica estatística multivariada: a Regressão Logística.

Adicionalmente, estima-se a taxa de recuperação do crédito, para clientes incumpridores, recorrendo à Regressão Beta, com base no histórico do crédito de cada cliente.

Neste trabalho propõe-se, ainda, um modelo para a estimação do *spread* a aplicar a um novo cliente assumido pela instituição bancária, em função da probabilidade de *default* (incumprimento) e da taxa de recuperação estimada.

Palavras-chave: Risco de Crédito, Regressão Logística, Regressão Beta, Probabilidade de Default, Taxa de Recuperação, Spread

Abstract

The analysis of credit risk in banking institutions and the measurement of risk is of extreme importance to institutions, since the granting of credit is their main activity. The ability to distiguish "good" and "bad" clients, is a decisive process in the constitution of the credit and therefore Credit Scoring models are implemented, quantitive models that consist of a statistical analysis of the credit quality.

The puporse of this dissertation is to estimate the probability of default of each client depending on the socio-economic and demographic variables, taking the database of a consumer credit portfolio of a Cape Verde bnk, over a multivariate statistical technique: Logistic Regreesion.

Additionally, a recovery rate of the credit is estimated, for defaulting clients using a Beta Regression, based on client history.

This survey also proposes a model for determining the spread to apply for a new client assumed by the banking institution, and the spread is a function of the probability os default and of the recovery rate.

Keywords: Credit Risk, Logistic Regression, Beta Regression, Probability of Default, Recovery Rate, Spread

Conteúdo

1 Introdução **1**

2 Risco de Crédito e Spread **3**

2.1 Modelos de análise de Risco de Crédito 3

2.1.1 Risco de Crédito 3

2.1.2 Modelos de Risco de Crédito 4

2.1.3 Modelos de Classificação de Risco 5

2.2 Estimação do *Spread* 10

2.2.1 *Spread* 10

2.2.2 Medidas de Risco de Crédito 11

2.2.3 Estimação do *Spread* - uma proposta de modelação 11

3 Modelos Lineares Generalizados **15**

3.1 A Família Exponencial 15

3.2 Média e Variância 16

3.3 As componentes dos Modelos Lineares Generalizados 17

3.4 Metologia dos Modelos Lineares Generalizados 18

3.5 Estimação dos parâmetros 18

3.5.1 Método dos Scores de Fisher 19

3.6 Testes de Hipóteses sobre $\widehat{\beta}$ 20

3.7 Modelo de Regressão Logística 22

3.8 Modelo de Regressão Beta 25

3.9 Qualidade de Ajustamento 26

3.9.1 Função Desvio 27

3.10 Análise de Resíduos 29

3.11 Selecção dos Modelos 30

3.11.1 Método Stepwise para a escolha das covariáveis 31

3.11.2 Critério de Informação Akaike 31

3.12 Observações discordantes 32

4 Modelação de Risco de Crédito - Aplicação **35**
4.1 A Carteira de Crédito 35
4.1.1 Base de dados 35
4.1.2 Definição de Cliente incumpridor 37
4.1.3 Análise Estatística das Variáveis 39
4.2 Probabilidade de *Default* - Regressão Logística 45
4.2.1 Ajustamento dos dados - Probabilidade de *Default* 45
4.2.2 Análise dos resíduos 50
4.2.3 Estimação da probabilidade de *Default* 54
4.3 Proporção das Prestações Pagas - Regressão Beta 57
4.3.1 Ajustamento dos dados - Proporção de Prestações Pagas 58
4.3.2 Análise de Resíduos 62
4.3.3 Estimação da Proporção das Prestações Pagas 65
4.4 Taxa de Recuperação do Crédito - Regressão Beta 67
4.4.1 Ajustamento de dados - Taxa de Recuperação 67
4.4.2 Análise de Resíduos 69
4.4.3 Estimação da Taxa de Recuperação 70
4.5 Estimação do *Spread* - Metodologia Actuarial 72

5 Conclusão **75**

Lista de Figuras

4.1 Histogramas Variáveis Qualitativas 40
4.2 Caixa-e-Bigodes: Variável *Valor Empréstimo* 41
4.3 Caixa-e-Bigodes: *Prazo* e *Prestações Pagas* 41
4.4 Histogramas Variáveis Quantitativas 42
4.5 1° Critério: Variável *default vs* variáveis Quantitativas 43
4.6 1° Critério: Relação entre a variável *default* e as variáveis Quantitativas . . 43
4.7 2° Critério: Relação entre a variável *default* e as variáveis Quantitativas . . 44
4.8 2° Critério: Variável *default vs* variáveis Quantitativas 45
4.9 Prob. *default*: Desvios residuais reduzidos - 1° Critério 51
4.10 Prob. *default*: Observações com repercussão Elevada - 1° Critério 52
4.11 Prob. *default*: Distâncias de Cook - 1° Critério 52
4.12 Prob. *default*: Desvios residuais reduzidos - 2° Critério 53
4.13 Prob. *default*: Observações com repercussão Elevada - 2° Critério 53
4.14 Prob. *default*: Distâncias de Cook - 2° Critério 54
4.15 Distribuição da probabilidade de *default* da carteria 56
4.16 PPagas: Análise de Resíduos - 1° Critério 63
4.17 PPagas: Distâncias de Cook - 1° Critério 63
4.18 PPagas: Análise de Resíduos - 2° Critério 64
4.19 PPagas: Distâncias de Cook - 2° Critério 64
4.20 Distribuição da *Proporção das Prestações Pagas* da carteria 67
4.21 Taxa de Recuperação - Análise de Resíduos 69
4.22 Taxa de Recuperação - Distâncias de Cook 70
4.23 Distrituição da *Taxa de Recuperação* da carteira 72

Lista de Tabelas

2.1 *Rating* da agência *Standard & Poor's* 9

4.1 Definição das Variáveis 36
4.2 Descrição das Categorias 37
4.3 Definição de Cliente incumpridor 38
4.4 Definição de Cliente incumpridor 38
4.5 Análise Preliminar das Variáveis 39
4.6 Prob. *default*: Modelo Completo - 1° Critério 46
4.7 Prob. *default*: Modelo de ajustamento final - 1° Critério 48
4.8 Prob.*Default*: Modelo Completo - 2° Critério 49
4.9 Prob. *default*: Modelo de ajustamento final - 2° Critério 50
4.10 Prob. *default*: Cliente Padrão 55
4.11 Prob. *default*: Ajustamento da Probabilidade de *default* 55
4.12 Clientes Ilustrativos 56
4.13 Prob. *default*: Exemplos 56
4.14 Prob. *default*: *Odds Ratio* 57
4.15 PPagas: Modelo Completo - 1° Critério 59
4.16 PPagas: Modelo final de ajustamento - 1° Critério 60
4.17 PPagas: Modelo Completo - 2° Critério 61
4.18 PPagas: Modelo final de ajustamento - 2° Critério 62
4.19 PPagas: Cliente Padrão 65
4.20 PPagas: Ajustamento da *Proporção das Prestações Pagas* 66
4.21 Proporção em Dívida - Modelo Completo 68
4.22 Proporção em Dívida - Modelo de ajustamento final 69
4.23 Modelo de regressão para LGD: Cliente Padrão 71
4.24 Modelo de regressão para *LGD* 71
4.25 Exemplos - Estimação do *spread* 73

1

Introdução

Nas últimas décadas, devido à crise financeira, têm vido a ocorrer profundas mudanças no seio das instituições bancárias, pelo que se tornou fundamental para estas controlar o risco de crédito. Muitas destas mudanças foram originadas pela competitividade dos mercados, pela expansão dos mercados de capitais, pelas alterações desregulares da taxa de juro, por *spreads* altos ou pelo aumento da probabilidade de incumprimento. E acresceu a necessidade de controlar e de gerir eficazmente o risco de modo a aferir a probabilidade de incumprimento na concessão de um crédito.

Em Cabo Verde, em comparação com outras economias similares, o diferencial das taxas de juro ou spread das instituições bancárias, é geralmente, alto, o que condiciona de certa forma o seu desenvolvimento no sector. E a análise de risco é um processo essencialmente intuitivo, baseado na experiência dos analistas de crédito, pelo que, perante a crescente pressão para a maximização das receitas das instituições, estas foram levadas a procurar mecanismos mais eficientes para seleccionar novos clientes com baixo perfil de risco e ao mesmo tempo controlar e minimizar as perdas. O aparecimento de novas tecnologias, o aumento da procura por crédito, bem como por uma questão de qualidade de serviço, a necessidade de responder o mais rápido possível às solicitações levou ao desenvolvimento e aplicação de sofisticados modelos estatísticos na gestão de risco de crédito, designados por *Credit Scoring*.

Assim, o objectivo desta dissertação, com base na carteira de crédito ao consumo de uma Instituição Bancária de Cabo Verde, passa por estimar a probabilidade de *default* (incumprimento) da carteira e do cliente através da Regressão Logística, em função de variáveis sócio-económicas e demográficas. O objectivo final passa por calcular o *spread*

de um cliente novo com intenção de constituir um crédito, bem como o *spread* da carteira, aplicando um modelo em tempo discreto, sendo necessário estimar, também, a taxa de recuperação do cliente, como componente necessária ao cálculo do *spread*, utilizando a Regressão Beta. Resultados prévios haviam sido obtidos por [Fer12], numa versão preliminar da base de dados recolhida.

Um breve resumo da estruturação deste trabalho, no capítulo 2 são apresentados os modelos de crédito de risco, como o seu conceito e as suas diversas características. É, também, descrito uma proposta de modelação para o *spread*, um modelo a tempo discreto, do qual originou um artigo [EGFS14]. No capítulo 3 é apresentado um alargado resumo sobre os Modelos Lineares Generalizados, e em particular, e com mais detalhe, a Regressão Logística e a Regressão Beta, com o objectivo de serem aplicados na estimação da probabilidade de *default* e da taxa de recuperação. Por fim, no capítulo 4 é onde se encontra a fundamentação prática deste trabalho e consiste na exposição dos resultados obtidos para a estimação da probabilidade de *default* através de uma técnica estatística multivariada, a Regressão Logística; para a estimação da taxa de recuperação, tendo sido utilizado a Regressão Beta e para estimação do *spread* em função das últimas componentes referidas, através de um modelo proposto. Como análise complementar, estimou-se a proporção de prestações pagas, tendo por objectivo estimar a percentagem de empréstimo que se encontrará pago no vencimneto do contrato de crédito.

Esta dissertação, na persperctiva do risco de crédito, tem a sua importância, uma vez que é proposto um modelo de estimação do *spread* através de uma metodologia actuarial e que teve como resultado a construção de um artigo ([EGFS14]). E contribuiu-se com uma análise detalhada de uma carteira de crédito ao consumo de uma Instituição Bancária de Cabo Verde, tendo-se avaliado questões importantes como: a estimação da probabilidade de incumprimento, a estimação da taxa de recuperação e como consequência a estimação do *spread* em função das estimações anteriores.

2

Risco de Crédito e Spread

A concessão de crédito é uma das principais componentes da actividade das instituições bancárias e de algumas instituições financeiras, pelo que é fundamental que as entidades analisem as propostas, adoptando procedimentos que lhes permitam, eficaz e eficientemente aferir o risco dos créditos e melhorar a forma de corrigir o surgimento de acontecimentos negativos para as instituições.

No âmbito desta dissertação pretende-se apresentar um modelo que sirva de base à análise de risco de um cliente de crédito, com base nas características do cliente e do crédito solicitado, bem como identificar a probabilidade de *default* e a taxa de recuperação do crédito concedido. Estas duas medidas permitirão definir um *spread* adequado que permita ao credor a compensação do risco de incumprimento do contrato.

2.1 Modelos de análise de Risco de Crédito

2.1.1 Risco de Crédito

O Risco de Crédito é algo que está presente no quotidiano de qualquer instituição, seja uma empresa da área financeira, como uma empresa de serviço comercial ou industrial e define-se como a possibilidade de perdas resultantes do não recebimento de valores contratados junto a clientes. Para determinar o risco de crédito de um cliente, com maior ou menor exactidão, pode-se proceder a avaliações do risco.

A palavra *crédito* deriva do latim *creditum* e significa confiança ou segurança de alguma coisa e para uma instituição bancária ou financeira, refere-se principalmente à actividade de colocar um valor à disposição de um tomador sob a forma de um empréstimo ou financiamento, mediante compromisso de pagamento do valor constituído por empréstimo numa data futura.

O risco de crédito resulta da possibilidade de perdas resultantes pelo não recebimento de valores contratados junto a clientes, ou seja, risco de crédito pode ser definido pelas perdas geradas por um evento de *default* do tomador ou pela decadência da sua qualidade de crédito e entenda-se por *default*, a incapacidade para cumprir as condições de uma obrigação, resultando daí uma dívida do devedor perante o credor, a instituição que constituiu o contrato de crédito.

O risco de crédito pode-se dividir entre três componentes, o risco de *default*, o risco de exposição e o risco de recuperação. O risco de *default* está associado à probabilidade de ocorrer um acontecimento de *default*, isto é, de incumprimento por parte do tomador num certo período de tempo. O risco de exposição deriva da incerteza em relação ao valor de crédito no momento de *default*. O risco de recuperação refere-se à incerteza quanto ao valor que pode ser recuperado pelo credor no caso de incumprimento do cliente e este depende do tipo de *default* ocorrido e das características do processo de crédito, como valor, prazo e garantias. O risco de *default* é também designado por "risco cliente", pois está vinculado às características intrínsecas do tomador de crédito, os riscos de exposição e de recuperação são nomeados de "risco operação", uma vez que estão associados a factores específicos do crédito, ver [And04].

A mensuração do risco de crédito é um processo essencial para as instituições bancárias ou financeiras, uma vez que quantifica a possibilidade da instituição vir a sofrer perdas em processos de crédito e o risco de *default* constitui a principal variável desse processo. De forma a estudar esta variável foram desenvolvidos os *Modelos de Risco de Crédito*.

2.1.2 Modelos de Risco de Crédito

Os Modelos de Risco de Crédito são ferramentas e aplicações que têm por objectivo principal mensurar o risco de tomadores ou de uma carteira de crédito como um todo. Segundo [And04], os modelos de risco podem ser classificados em três grupos: os modelos de classificação de risco, os modelos estocásticos de risco de crédito e os modelos de risco de carteira.

Os modelos de classificação de risco são modelos que avaliam o risco de um tomador ou de um crédito, atribuíndo uma medida que representa a probabilidade de ocorrência de *default* e geralmente é expressa na forma de uma classificação de risco - *rating* - ou de pontuação - *score*. Os modelos estocásticos de risco de crédito têm por objectivo avaliar o comportamento estocástico, não determinístico, do risco de crédito ou das variáveis que o determinam. O modelo de risco de carteira visa estimar a distribuição estatística das perdas ou de valor de uma carteira de crédito, a partir da qual são estraidas medidas que quantificam o risco da carteira.

Cada uma das categorias acima descritas possui diferentes objectivos em relação ao que se pretende prever ou modelar. Nos modelos de classificação de risco de crédito, o fenómeno que se pretende modelar é a ocorrência ou não de um evento de incumprimento. Os modelos estocásticos têm o seu foco na modelação do comportamento de variáveis relacionadas com o evento de *default* de um devedor. Já nos modelos de carteira, o objectivo é modelar a distribuição de perdas na carteira.

Neste capítulo, ir-se-á apresentar com mais detalhe os modelos de classificação de risco.

2.1.3 Modelos de Classificação de Risco

Os modelos de risco de classificação têm como objectivo analisar o crédito de forma a auxiliar o credor na tomada de uma decisão a partir da avaliação de diversas informações sobre o tomador de crédito, originando uma avaliação do risco. Uma boa gestão de risco de crédito por parte das instituições financeiras é indispensável, para que se evite a insolvência das mesmas, uma vez que a concessão de crédito constitui, como já referido, uma das suas principais actividades.

A análise de crédito pode ser tratada tendo em conta duas metodologias: a qualitativa e a quantitativa. A análise quantitativa utiliza informação proveniente de modelos estatísticos e econométricos que permitem uma mensuração do risco do tomador de crédito, através de Modelos de *Scoring* e de *Rating*. A análise qualitativa remete para julgamentos subjectivos por parte do analista de crédito, em relação à capacidade de pagamento do tomador de crédito, designados por Modelos Especialistas.

Por definição, os Modelos Especialistas, envolvem decisões individuais quanto à decisão de conceder ou não o crédito, segundo um conjunto de regras. Neste processo, a decisão baseia-se na experiência na área, na disponibilidade de informações e na sensibilidade de cada analista quanto ao risco do negócio. As informações que são necessárias para a análise subjetiva da capacidade financeira dos clientes são tradicionalmente conhecidas como os C's do crédito: Caráter - intenção de um cliente pagar a sua dívida

-, Capacidade - habilidade de um cliente em honrar os seus compromissos -, Capital - situação financeira em termos de decomposição, aplicação e financiamento -, Colateral - garantias que podem ser oferecidas pelo cliente - e Condições - Sensibilidade da capacidade de pagamento em função dos fatores externos. [Sec02] propõe mais um C, o Conglomerado - informações referentes à situação de empresas do mesmo grupo económico -, que actualmente é considerado nas análises de crédito das intituições financeiras.

A principal vantagem da abordagem qualitativa é a especifidade com que é tratado cada caso, a principal desvastagem é a sua dependência na experiência do avaliador, o baixo volume de produção e o envolvimento pessoal do concedente do crédito. Por outro lado, na análise quantitativa, as regras são bem definidas em relação às características dos clientes e às operações de crédito, e são baseadas, em geral, em modelos estatísticos.

2.1.3.1 Modelos de *Credit Scoring*

Os modelos de *Credit Scoring* são normalmente utilizados para avaliação de um cliente que pretende constituir um crédito, a partir de características do proponente e de informações sobre o próprio crédito, como o seu valor, prazo, garantias, por exemplo. Os modelos são baseados em técnicas de análise estatística e geram uma pontuação (*score*) que representa a propensão de risco associada ao tomador de crédito. Embora a medida de risco seja normalmente fornecida numa escala contínua, esta pode ser categorizada para originar uma medida ordinal.

Na extensa literatura sobre risco de crédito existem várias definições de *Credit Scoring*. Por exemplo, [Lew92] define *Credit Scoring* como um processo em que a informação sobre o solicitante é convertida em números que, de forma combinada, forma um *score*, que representa o perfil de risco do solicitante. [Mes97] acrescenta que *Credit Scoring* é um método estatístico utilizado para prever a probabilidade de um solicitante entrar em incumprimento. Usando dados históricos, o *Credit Scoring* isola as características dos clientes que entraram em situação de *default*, produzindo, então, um *score* que a instituição utiliza para classificar o candidato ao crédito em termos de risco e para decidir quanto à aprovação do crédito.

Para [CAN98], os modelos tradicionais de *Credit Scoring* atribuem pesos, determinados estatísticamente, de modo a que se possa criar um *score* de crédito. E este é representivo do risco de perda. Segundo [ACn], pode ser estabelecida uma pontução máxima para a qual é aceite o crédito, de modo a que se possa comparar o *score* de um novo cliente. O objectivo é pré-identificar factores chave que determinem a probabilidade de *default*, de modo a que a sua combinação ou ponderação possa produzir uma pontuação quantitativa que auxilie na avaliação do risco.

Segundo [Lew92], o primeiro modelo estatístico de análise de crédito foi desenvolvido em meados de 1945. Os primeiros modelos destinavam-se ao crédito ao consumo e o uso dos modelos foi expandido devido à massificação do mercado de crédito, o que obrigou os analistas a uma maior rapidez e homogeneidade no tratamento dos seus clientes. Por outro lado, a evolução dos sistemas informáticos possibilitou o tratamento estatístico adequado a esse aumento de dados. Embora o uso de métodos de *Credit Scoring* seja direccionado para a decisão de conceder ou não o crédito, algumas instituições também os utilizam para determinar o montante de crédito a ser concedido, como refere [CAN98].

Em resumo, a metodologia básica para o desenvolvimento de um modelo de *Credit Scoring*, segundo [SA02], deve ter em conta as seguintes etapas: planeamento e definições, os mercados e produtos de crédito para os quais o sistema será desenvolvido, bem como a definição de cliente incumpridor; identificação dos factores, caracterização do candidato ao crédito e selecção das variáveis significativas para o modelo; planeamento amostral e colecta de dados; determinação da fórmula de classificação através de técnicas estatísticas; determinação do ponto de corte a partir do qual o cliente é classificado como cumpridor ou bom pagador, ou seja, é o ponto a partir do qual a instituição financeira pode aprovar a concessão do crédito.

Tipos de *Credit Scoring*

Segundo [ACn], os modelos de *Credit Scoring* são divididos em duas categorias: *Modelos de Aprovação de Crédito* (*Credit Scoring* propriamente dito) e os de *Modelos de Classificação Comportamental*, também conhecidos como *Behavioural Scoring*.

Os modelos de *Credit Scoring* propriamente ditos, são ferramentas que dão suporte à avaliação da capacidade de crédito para novos clientes, sendo o principal objectivo estimar a probabilidade de um novo requerente de crédito se tornar incumpridor num determinado período.

O modelo *Behavioural Scoring* é uma ferramenta que tem em consideração os aspectos comportamentais e as actividades dos clientes existentes na instituição e prevê eventos associados ao risco de crédito, como o incumprimento e os pagamentos em dia, entre outras características. E tem como objectivo estimar a probabilidade de incumprimento de um cliente que já possuiu um produto ou um crédito com a mesma instituição financeira.

Os modelos de aprovação de crédito destinam-se essencialmente à concessão e volume de crédito, os modelos de classificação comportamental são usados para gestão dos

limites de crédito, cobrança preventiva e outras estratégias.

Os modelos de *Credit Scoring* são baseados em técnicas de análise estatística multivariada como modelos de Regressão linear, Regressão Logística ou em modelos de inteligência artificial como redes neuronais. Nesta dissertação ir-se-á utilizar a formulação mais comum dos modelos, a probabilidade *default* será obtida através de um modelo de Regressão Logística.

Vantagens e Desvantagens dos modelos de *Credit Scoring*

Segundo [ACn] as principais vantagens dos modelos *Credit Scoring* são:

- ***Consistência***: são modelos bem elaborados que utilizam a experiência da instituição e ajudam a administrar objectivamente os créditos dos clientes já existentes e dos novos requerentes;
- ***Facilidade***: os modelos de *Credit Scoring* buscam a simplicidade e a fácil intrepretação, com instalação relativamente fácil;
- ***Melhor organização da informação de crédito***: a sistematização e organização das informações contribuem para a melhoria do processo de concessão de crédito;
- ***Redução metodologia subjectiva***: a utilização do método quantitativo com regras claras e bem definidas contribui para a diminuição da subjectividade na avaliação do risco de crédito;
- ***Maior eficiência do processo***: aumenta a qualidade do serviço prestado ao cliente, trazendo redução de tempo e maior eficiência a este processo.

[Sem09] e [ACn] enunciam as principais desvantagens dos modelos de *Credit Scoring*:

- ***Custo de desenvolvimento***: desenvolver um sistema de *Credit Scoring* acarreta custo, não somente com a instalação do sistema, mas também com o suporte para a sua construção, como por exempo, profissionais capacitados e equipamentos;
- ***Escassez e qualidade dos dados***: os modelos, normalmente, são desenvolvidos com base nas observações presentes nas bases de dados das instituições, em que a qualidade nem sempre é salvaguardada;
- ***Excesso de confiança nos modelos***: algumas estatísticas podem estimar por valores superiores a eficácia dos modelos, provocando com que alguns analistas, principalmente os menos experientes, considerem-nos perfeitos sem questionar os seus resultados;
- ***Interpretação equivocada das classificações***: um sistema é complexo t, e eventuais erros no desenvolvimento do modelo de *Credit Scoring*, podem acarretar custos para a instituição ou resultar em situações danosas na concessão do crédito.

2.1.3.2 Modelos de *Rating*

Os modelos de *Credit Rating*, segundo [And04], são modelos que utilizam um sistema de mensuração de risco de crédito baseado em pontuação - *rating* - e enquadram os riscos em "classes de risco", previamente definidas. Às "classes de risco" são atribuídas notas que refletem diferentes graus de risco, de acordo com uma escala pré-determinada, que é parte integrante do modelo de avaliação. A definição da escala resulta a partir de opiniões técnicas sobre a capacidade futura, a responsabilidade jurídica e a vontade de um devedor efectuar, dentro do prazo, o pagamento das obrigações por ele contraídas. Logo, os *ratings* de crédito são uma opinião prospectiva sobre a qualidade de crédito.

Actualmente, as instituições financeiras desenvolvem internamente os seus próprios modelos de *Credit Rating* ou utilizam os que são facultados por agências de r*ating*, organizações que se especializam em avaliar o risco de crédito, como por exemplo *Standard & Poor's* e *Fitch Ratings*. Cada agência aplica a sua própria metodologia para medir a qualidade de crédito e usa uma escala de *ratings* específica para publicar opiniões de *ratings*. Normalmente, os *ratings* são expressos por meio de letras que variam, por exemplo, de 'AAA' a 'D', para comunicar a opinião da agência sobre o nível relativo de risco de crédito. A Tabela 4.1 representa um exemplo de *rating* da agência de *rating Standard & Poor's*.

Classificação	Significado
AAA	Capacidade extremamente forte para honrar compromissos financeiros; Rating mais alto
AA	Capacidade muito forte para honrar compromissos financeiros
A	Forte capacidade para honrar compromissos financeiros, porém é de alguma forma suscetível a condições económicas adversas
BBB	Capacidade adequada para honrar compromissos financeiros, porém mais sujeito a condições económicas adversas
BBB-	Considerado o nível mais baixo da categoria de grau de investimento pelos participantes do mercado
BB+	Considerado o nível mais alto da categoria de grau especulativo pelos participantes do mercado
BB	Menos vulnerável no curto prazo, porém enfrenta atualmente grande suscetibilidade a condições adversas de negócios
B	Mais vulnerável a condições adversas de negócios, porém atualmente apresenta capacidade para honrar compromissos financeiros
CCC	Atualmente vulnerável e dependente de condições favoráveis de negócios para honrar seus compromissos financeiros
CC	Atualmente fortemente vulnerável
C	Um pedido de falência foi registrado ou acção similar, porém os pagamentos das obrigações financeiras continuam sendo realizados
D	Inadimplente em seus compromissos financeiros.

Tabela 2.1: *Rating* da agência *Standard & Poor's*

Uma vez que eventos e desenvolvimentos futuros não são previsíveis, a atribuição de um *rating* de crédito não é uma ciência exacta, por exemplo um crédito cujo *rating* é "AA" considerado pela agência de *rating* como tendo uma qualidade de risco inferior do que um crédito com o *rating* "BBB", o rating "AA" não é uma garantia de que não haverá

ocorrência de *default*, apenas que esta é menos provável no primeiro caso do que no segundo.

As classificações de *rating* não são fixas, são revistas regularmente e existem várias razões que levam a um ajuste nos *ratings*. Podem estar relacionadas com as alterações gerais no ambiente económico ou de negócios, ou estarem mais estreitamente relacionadas com circunstâncias particulares que afetam uma indústria específica, entidade ou título de dívida individual.

Vantagens e Desvantagens dos modelos de *Credit Rating*

Segundo [Fin03], o sistema de *rating* de risco de crédito desenvolvido pelas agências de *rating* possui uma vantagem evidente, facilita uma visão mais abrangente do mercado, por incorporar um universo de empresas e análises. No caso de sistemas próprios dos bancos e grandes empresas, a enorme vantagem está na sistematização do processo de interpretação dos riscos, uma vez que o modelo já tem a definição básica dos riscos a serem identificados e traz as respectivas pontuações já pré-definidas, tornando as avaliações mais homogéneas. E a desvantangem do sistema de *rating* de risco de crédito é que este está sujeito a variações qualitativas, influenciadas pela competência técnica e experiência dos avaliadores, pela metodologia de mensuração de riscos empregada, pelo modelo de coleta, análise e avaliação de dados; uniformidade e consistência de aplicação da metodologia, qualidade e confiabilidade das fontes de informações utilizadas no desenvolvimento da análise. É evidente que a qualidade e confiabilidade das fontes de informações é, de entre os itens acima referidos, o mais importante para a validação de qualquer sistema de *rating*, sem ele, o sistema estará seriamente prejudicado.

2.2 Estimação do *Spread*

2.2.1 *Spread*

O *spread* define-se pela diferença entre o preço de compra e o preço de venda, aplicado pelas instituições financeiras, numa transação monetária como a transação de um título. Por outro lado, o chamado *spread* bancário é um valor percentual definido pela diferença entre a taxa de juro que as instituições financeiras pagam na aquisição do dinheiro e a que cobram aos clientes. É também conhecida como "taxa de risco".

Para as instituições bancárias ou financeiras, o *spread* define-se como a medida de risco de crédito que um determinado cliente representa e o seu valor provém da análise de vários factores do cliente e do empréstimo em causa. Mas, mais especificamente, quanto menor é o risco para a instituição menor será o *spread*, reduzindo assim o custo

do empréstimo para o cliente.

Para definir um *spread* adequado para um cliente, que permita ao credor a compensação do risco de incumprimento do contrato, é necessário identificar a probabilidade de *default* e a taxa de recuperação do crédito concedido.

2.2.2 Medidas de Risco de Crédito

O risco de crédito é um dos riscos mais comuns numa instituição financeira ou bancária, uma vez que a concessão de crédito é a sua maior actividade e o risco de crédito é o risco de perda devido a uma falta de pagamento por parte do tomador de crédito.

Desta forma, é necessário analisar e quantificar o risco de crédito, com intenção de identificar o nível de risco presente numa operação de crédito. Para avaliar o incumprimento de clientes utilizam-se essencialmente os indicadores seguintes:

- Probabilidade de *Default* (PD) : probabilidade de um cliente entrar em incumprimento num dado horizonte temporal;
- Processos de Exposição (*Exposure at Default* - EAD) : valor em dívida pelo cliente, à instituição, no momento do incumprimento;
- Taxa de Recuperação (*Recovery Rate* - R) : é a percentagem do montante de crédito concedido que a instituição financeira recupera, em caso de ocorrer *default*;
- *Loss Given Default* (LGD) : valor que a instituição perde efectivamente, quando um cliente entra em incumprimento e pode ser definida, também, através da taxa de recuperação, $R = 1 - L$.

2.2.3 Estimação do *Spread* - uma proposta de modelação

Na gestão de risco de crédito um dos processos mais importantes é a definição adequada do *spread* a aplicar num contrato de crédito. Pelo que o objectivo desta secção é propor um modelo de estimação do *spread* em função da taxa de recuperação e da probabilidade de *default*.

Primeiramente irá-se-á apresentar as definições das variáveis e parâmetros necessários para a construção do modelo de estimação do *spread*.

A evolução dos *cash-flows* de um crédito pode ser descrito como um processo estocástico $(X_t)_{t\in\varphi}$, com o conjunto de tempo φ pertecente a um subconjunto dos números inteiros. Em função da complexidade da evolução dos *cash-flows*, pode-se recorrer à modelação deste fenómeno através de um processo de Markov ou até mesmo através de

uma martingala, como é usual na literatura, ver, por exemplo [MFE05].

Para efeitos de modelação dos *cash-flows* dos clientes incumpridores, o tempo de *default* (incumprimento) τ deve ser, pelo menos, uma variável aleatória. Donde é necessário que τ seja um tempo de paragem e associada a esta variável ter-se-á a probabilidade de incumprimento, pelo que deverá ser um parâmetro de interesse do modelo. Note-se que, uma hipótese natural, seria a de que τ corresponde ao tempo de paragem relativamente à filtração natural do processo $(X_n)_{n\in\{0,1,\ldots,T\}}$, ou seja para $\{\tau \geq n\} \in \mathcal{A}_n$, com $\mathcal{A}_n$ sigma-àlgebra gerada por X_k, para $0 \leq k \leq n$, onde o tempo de início de incumprimento τ é definido perfeitamente pelas variáveis aletórias X_k da carteira até ao tempo presente para qualquer n. No entanto, devido à natureza do modelo que se irá propor, esta hipótese não será necessária.

A recuperação dos *cash-flows* deve ser também representada por um processo de Markov. No entanto, informação sobre a recuperação nem sempre é fiável, existem geralmente dúvidas sobre a sua qualidade, pelo que neste processo ir-se-á considerar como um parâmetro constante λ.

Por último, usualmente, considera-se, que o *spread* deverá ser um processo de Markov, como se pode ver em [MFE05]. Teoricamente, o *spread* de crédito $s(t,T)$, no tempo t e com maturidade T, para uma obrigação de cupão zero com possibilidade de incumprimento (*defaultable zero coupon bond*) com preço no tempo t é dado por $p_1(t,T)$ e é tal que:

$$p_1(t,T)(1+s(t,T))^{T-t} = p_0(t,T) \tag{2.1}$$

sendo $p_0(t,T)$ o preço no tempo t de uma obrigação de cupão zero livre de incuprimento (*default-free zero coupon bond*) com vencimento em T. A partir de (2.1) pode ser obtida uma expressão que descreve o *spread*:

$$s(t,T) = \sqrt[T-t]{\frac{p_1(t,T)}{p_0(t,T)}} - 1. \tag{2.2}$$

Como referido, o propósito do modelo que se irá propor, é determinar o *spread* em função da taxa de recuperação e da probabilidade de default, pelo que, no que se segue, se apresenta uma metodologia actuarial para a modelação do *spread*.

Modelo de uma carteira em tempo discreto - Metodologia actuarial

Seja $(X_n)_{n\in\{0,1,\ldots,T\}}$ um processo estocástico que descreve os valores das obrigações da carteira para um credor. Considere-se $(\Omega, \mathcal{A}, \mathbb{P})$ um espaço de probabilidade em que as variáveis aleatórias X_n, para $n \in \{0,1,\ldots,T\}$, são definidas para cada $\omega \in \Omega$, donde $X_n(\omega)$ é o valor, para o credor, da obrigação do cliente ω na data $n \in \{0,1,\ldots,T\}$.

Suponha-se que as obrigações na carteira, representadas por $(X_n)_{n\in\{0,1,\ldots,T\}}$ não estão sujeitas a incumprimento, tendo-se que $X_0 \equiv 0$ e que X_T representa o valor total do empréstimo concedido..

Note-se que, o incumprimento pode ocorrer num tempo aleatório τ, que se designa como tempo de incumprimento da carteira, de modo que para cada $n \in \{0, 1, \ldots, T\}$, se tem $\{\tau \geq n\} \in \mathcal{A}$.

De acordo com [EGFS14], o modelo assume os seguintes pressupostos:

1. Existe uma função $F : \mathbb{R} \times \Omega \mapsto \mathbb{R}$ e um parâmetro $\lambda \in [0, 1]$, designado de taxa de recuperação da carteira que, para um tempo de maturidade T, se tem:

$$\mathbb{E}\left[F(X_T, \cdot)\right] = \lambda \mathbb{E}\left[X_T\right]. \tag{2.3}$$

2. Seja $(\widetilde{X}_n)_{n\in\{0,1,\ldots,T\}}$ um conjunto de variáveis aleatórias que denotam um processo estocástico descrito para os valores das obrigações, as quais estão agora sujeitas a incumprimento. Desta forma, para a função F, no tempo de idade T, ter-se-á:

$$\widetilde{X}_T = X_T \mathbb{I}_{\{\tau > T\}} + F(X_T, \cdot)\mathbb{I}_{\{\tau \leq T\}}.$$

   Note-se que, se o evento de ocorrência de *default* não ocorrer então, $\widetilde{X}_T = X_T$ e se o default ocorrer antes do vencimento do contrato ter-se-á que $\widetilde{X}_T = F(X_T, \cdot)$.

3. A taxa de juro de risco r e o *spread*, calculado, na maturidade T, e denotado por s_T, são ambos constantes.

4. O processo da carteira de obrigações $(X_n)_{n\in\{0,1,\ldots,T\}}$ e o tempo de incumprimento τ da carteira são independentes.

Com este conjunto de hipóteses, pelo princípio do valor esperado da metodologia actuarial, mostra-se que o *spread*, s_T, é uma função do incuprimento e da taxa de recuperação, como se pode ver em [Fer12].

Teorema 1: No âmbito do princípio do valor esperado da metodologia actuarial, tem-se que,

$$\mathbb{E}\left[X_T(1+r)^{-T}\right] = \mathbb{E}\left[\widetilde{X}_T(1+r)^{-T}(1+s_T)\right], \tag{2.4}$$

e se $\mathbb{E}\left[X_T\right] \neq 0$, o spread na data T é dado por:

$$s_T = \frac{(1-\lambda)\mathbb{P}\left[\tau \leq T\right]}{1-(1-\lambda)\mathbb{P}\left[\tau \leq T\right]}. \tag{2.5}$$

A demonstração do Teorema 1 é imediata recorrendo às propriedades do valor esperado.

Considere-se que Λ e Δ são variáveis aleatórias que representam a taxa de recuperação e a probabilidade de *default* de cada cliente da carteira, respectivamente, e que $\Lambda \cdot X_T = F(X_\tau, \cdot)$. Assim, a probabilidade de *default* da carteira é dada por:

$$\mathbb{E}[\Delta] = \mathbb{P}[\tau \leq T] \ ,$$

e a taxa de recuperação da carteira por:

$$\lambda = \mathbb{E}[\Lambda] \ .$$

Como é referido em [EGFS14], se Δ e Λ forem variáveis aleatórias independentes, pode-se definir o *spread* para cada cliente individual de uma carteira como:

$$s_{\text{cliente}} := \frac{(1-\Lambda)\,\Delta}{1-(1-\Lambda)\,\Delta}. \tag{2.6}$$

Note-se que, no caso de $(1-\Lambda)\,\Delta \ll 1$, tem-se a aproximação $s_{\text{cliente}} \approx (1-\Lambda)\,\Delta$ e se $(1-\lambda)\,\mathbb{P}[\tau \leq T] \ll 1$, pode-se considerar que $s_T \approx (1-\lambda)\mathbb{P}[\tau \leq T]$ e portanto, ter-se-á:

$$\mathbb{E}[s_{\text{cliente}}] \approx \mathbb{E}[(1-\Lambda)\,\Delta] = \mathbb{E}[1-\Lambda]\,\mathbb{E}[\Delta] \approx s_T. \tag{2.7}$$

Nesta secção, apresentou-se um modelo para a estimação do *spread* de uma carteira de crédito, através de uma metodologia actuarial, em função da probabilidade de *default* e da taxa de recuperação, e ainda, se estabelece uma ligação entre a formulação da estimação do *spread* dos clientes e da carteira, uma vez que se mostrou que é possível definir os *spreads* individuais de cada cliente de uma forma coerente.

3

Modelos Lineares Generalizados

Os Modelos Lineares e Generalizados (MLG), introduzidos por Nelder e Wedderburn em 1972, foram desenvolvidos com o objectivo de unificar modelos anteriormente desenvolvidos. Os autores mostraram que uma série de técnicas estatísticas, estudadas separadamente, podem ser formuladas de uma forma unificada, como uma classe de modelos de regressão. São casos particulares dos modelos lineares generalizados, por exemplo, o modelo linear de regressão linear clássico, o modelo de regressão logística, o modelo de regressão beta, entre outros.

Nesta dissertação, os Modelos Lineares Generalizados serão a base do estudo que se pretende realizar pelo que se fará uma exposição acerca deste tema, tendo como principal referência [TS00].

3.1 A Família Exponencial

Diz-se que uma variável aleatória Y tem distribuição pertencente à *Família Exponencial*, ver [MN89] se a sua função de densidade de probabilidade (f.d.p.) ou função de massa de probabilidade (f.m.p.) se puder escrever na forma:

$$f(y_i|\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{3.1}$$

onde θ e ϕ são parâmetros escalares e $a(\cdot)$, $b(\cdot)$ e $c(\cdot,\cdot)$ são funções reais conhecidas. São exemplos de distribuições da Família Exponencial as distribuições Normal, Gama, Binomial, Poisson, etc.

O parâmetro θ é designado por parâmetro de localização na forma canónica e ϕ, parâmetro estritamente positivo, é denominado por parâmetro de escala. Admite-se ainda que a função $b(\cdot)$ é diferenciável e que o suporte da distribuição não depende dos parâmetros. Desta forma, a Família Exponencial obedece às usuais condições de regularidade, ver [RS01].

3.2 Média e Variância

Seja Y uma variável aleatória pertencente à Família Exponencial com função de probabilidade definida como em (3.1). A função log-verosimilhança será dada por:

$$l(\theta_i;\phi,y_i) = ln[f(y_i|\theta_i,\phi)] = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi). \tag{3.2}$$

A função *Score* é definida por:

$$S(\theta_i) = \tfrac{\partial l(\theta_i;\phi,Y_i)}{\partial \theta_i},$$

tendo-se para a Família Exponencial,

$$S(\theta_i) = \frac{\theta_i - b'(\theta_i)}{a(\phi)} \tag{3.3}$$

bem como

$$\tfrac{\partial S(\theta_i)}{\partial \theta_i} = -\tfrac{b''(\theta_i)}{a(\phi)}$$

onde $b'(\theta_i)$ e $b''(\theta_i)$ correspondem à primeira e segunda derivada de $b(\theta_i)$, respectivamente.

Sob as condições de regularidade, ver [RS01], sabe-se que

$$\mathbb{E}[S(\theta_i)] = 0$$

e

$$\mathbb{E}[S^2(\theta_i)] = \mathbb{E}\left[\left(\tfrac{\partial l(\theta_i;\phi,Y_i)}{\partial \theta_i}\right)^2\right] = -\mathbb{E}\left[\tfrac{\partial^2 l(\theta_i;\phi,Y_i)}{\partial \theta_i^2}\right]$$

Desta forma, a partir das equações anteriores pode se estabelecer que:

$$\mu_i = \mathbb{E}[Y_i] = b'(\theta_i) \quad e \quad \mathbb{V}[Y_i] = a(\phi)b''(\theta_i). \tag{3.4}$$

A primeira equação de (3.4) permite verificar que o parâmetro canónico θ é função de μ, uma vez que

$$\theta_i = b'^{-1}(\mu_i). \tag{3.5}$$

A segunda equação de (3.4) permite concluir que Y é função do parâmetro canónico θ sendo, portanto, devido a (3.7), função do valor médio μ. Assim, a função $b''(\theta)$ expressa a relação entre a média e a variância e designa-se por *função de variância*, escrevendo-se

$$\mathbb{V}[\mu] = b''(\theta_i). \tag{3.6}$$

3.3 As componentes dos Modelos Lineares Generalizados

É comum referir-se que os Modelos Lineares Generalizados (MLG) são constituídos por três componentes:

- ***Componente Aleatória***

Esta componente do modelo estabelece que as variáveis aleatórias Y_i, que se pretendem modelar, são independentes com distribuição pertencente à Família Exponencial, em que

$$\mathbb{E}[Y_i|\boldsymbol{x}_i] = \mu_i = b'(\theta_i), \qquad i = 1, ..., n.$$

- ***Componente Estrutural ou Sistemática***

A componente sistemática dos MLG, também designada de *preditor linear*, consiste numa combinação linear das variáveis preditoras (ou covariáveis) dada por

$$\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

onde $\boldsymbol{x}_i$ é um vector de especificação de dimensão p tal que $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, ..., x_{i(p-1)})^T$ e $\boldsymbol{\beta}$ é um vector de parâmetros de dimensão p.

- ***Função de Ligação***

Outra característica destes modelos é a relação entre o valor esperado μ e o preditor linear η, que se estabelece através de

$$\mu_i = h(\eta_i) = h({\boldsymbol{x}_i}^T \boldsymbol{\beta})$$

onde $h(\cdot)$, designada por *função de ligação*, é uma função monótona e diferenciável, tal que $g(\cdot) = h^{-1}(\cdot)$.

Quando o preditor linear coincide com o parâmetro canónico, isto é, $\theta_i = \eta_i$, então a função de ligação denomina-se de *função de ligação canónica*.

3.4 Metologia dos Modelos Lineares Generalizados

Existem três fases que se devem seguir para modelar dados através dos Modelos Lineares Generalizados:

- Formulação dos modelos;
- Ajustamentos dos modelos;
- Selecção e validação dos modelos.

Numa primeira fase, a formulação do modelo, há a necessidade de examinar cuidadosamente os dados, para se determinar uma distribuição adequada que defina a variável resposta e que permita seleccionar as covariáveis que melhor explicitam o modelo em estudo. Deve-se ainda escolher a função de ligação, que depende do tipo de variável resposta e do estudo particular que se pretende efectuar.

A fase seguinte, o ajustamento do modelo, consiste na estimação dos parâmetros do modelo, isto é, na estimação do vector dos coeficientes β associados às covariáveis e respectivos erros padrão. Determinam-se intervalos de confiança e realizam-se testes de ajustamento, que permitam avaliar a qualidade do mesmo.

Numa última fase, procura-se encontrar submodelos que ainda se adequem aos dados, bem como procurar divergências que possam existir entre os dados e os valores preditos, localizar resíduos excessivos e possíveis *outliers* e/ou observações influentes.

3.5 Estimação dos parâmetros

Após a formulação do modelo que se considera adequado, há a necessidade de proceder à realização de inferências sobre esse modelo. Os Modelos Lineares Generalizados baseiam essa inferência na metodologia de máxima verosimilhança, ou seja, os parâmetros β que melhor explicitam os dados observados são estimados pelo método de máxima verosimilhança. No âmbito dos MLG, as equações de máxima verosimilhança não têm, regra geral, uma solução analítica, sendo necessário recorrer a métodos numéricos para a sua resolução.

Tendo em vista os Modelos Lineares Generalizados, [NW72] construiram um algoritmo para a resolução de tais equações, o que em muito contribuiu para o sucesso destes modelos, por se tratar de um algoritmo bastante geral, adaptável aos vários MLG e facilmente implementável de um ponto de vista computacional. Este algorítmo é designado de *Método Iterativo de Mínimos Quadrados Ponderados* e baseia-se no método dos Scores de Fisher, que se descreve na secção que se segue.

3.5.1 Método dos Scores de Fisher

Considere-se uma amostra de n observações e um Modelo Linear Generalizado definido por:

$$f(y_i|\theta_i,\phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\} \tag{3.7}$$

com função de ligação $h(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$.

Considerando a independência entre as variáveis Y_i, a função log-verosimilhança, em função de θ_i, será dada por:

$$l(\boldsymbol{\theta},\phi,\boldsymbol{y}) = \sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n}c(y_i,\phi). \tag{3.8}$$

Uma vez que $\mu_i = \mathbb{E}[\boldsymbol{Y}_i\boldsymbol{x}_i] = b'(\theta_i)$ e tendo em conta a função de ligação $h(\mu_i) = \eta_i = \sum_{j=1}^{p}x_{ij}\beta_j$ pode verificar-se que a função de log-verosimilhança é também uma função dos parâmetros de interesse $\boldsymbol{\beta}$.

Assim, os estimadores de máxima verosimilhança para $\boldsymbol{\beta}$ são obtidos como soluções do sistema de equações de verosimilhança

$$\frac{\partial l(\boldsymbol{\beta})}{\partial\beta_j} = \sum_{i=1}^{n}\frac{\partial l_i(\boldsymbol{\beta})}{\partial\beta_j} = 0, \quad j = 1,...,p. \tag{3.9}$$

A função Score, referida em (3.3), é obtida aplicando as regras de derivação da função composta sobre a equação anterior, donde se obtém o elemento genérico do Vector dos Scores,

$$s_j = \sum_{i=1}^{n}\frac{y_i - \mu_i}{\mathbb{V}[Y_i]}\frac{\partial\mu_i}{\partial\eta_i}x_{ij}, \quad j = 1,...,p \tag{3.10}$$

pelo que as equações de máxima verosimilhança para $\boldsymbol{\beta}$ serão dadas por:

$$\sum_{i=1}^{n}\frac{1}{\mathbb{V}[Y_i]}(y_i - \mu_i)x_{ij}\frac{\partial\mu_i}{\partial\eta_i} = 0, \quad j = 1,...,p \tag{3.11}$$

Como já referido anteriormente, as equações de máxima verosimilhança não são de fácil resolução, o que introduz a necessidade de resolução das mesmas através de métodos numéricos.

O *Método dos Scores de Fisher*, uma generalização do método de Newton-Raphson, introduz um algoritmo que permite a resolução de máxima verosimilhança para Modelos Lineares Generalizados e pode ser encontrado com máximo de detalhe em [TS00].

3.6 Testes de Hipóteses sobre $\widehat{\beta}$

Os testes de hipóteses sobre os parâmetros do modelo ajustado auxiliam na selecção das covariáveis que deverão ser incorporadas no modelo adoptado. Consoante o seu nível de significância, pode-se afirmar que um determinado parâmetro tem ou não influência no modelo.

Em relação a esta questão, vários cenários possíveis podem ser definidos:

- Hipótese de nulidade de um único parâmetro β_j, $j = 1, ..., p$:

$$H_0 : \beta_j = 0 \quad \textit{versus} \quad H_1 : \beta_j \neq 0. \tag{3.12}$$

  Esta hipótese corresponde a testar um submodelo com todas as covariáveis do modelo à excepção da covariável $\boldsymbol{x}_j$, relativamente ao parâmetro β_j.

- Hipótese de nulidade de vários parâmetros

$$H_0 : \boldsymbol{\beta}_r = 0 \quad \textit{versus} \quad H_1 : \boldsymbol{\beta}_r \neq 0. \tag{3.13}$$

  Esta hipótese corresponde a testar um modelo sem as r covariáveis relativas aos parâmetros supostos sob a hipótese H_0.

Em suma, estas hipóteses permitem testar a validade estatística de submodelos do modelo original e ser-nos-ão úteis na escolha do "melhor" modelo de ajustamento aos dados.

As hipóteses acima formuladas podem ser generalizadas por um teste da seguinte forma

$$H_0 : \boldsymbol{C\beta} = \boldsymbol{\xi} \quad \textit{versus} \quad H_1 : \boldsymbol{C\beta} \neq \boldsymbol{\xi}, \tag{3.14}$$

onde $\boldsymbol{C}$ é uma matriz $q \times p$, com $q \leq p$, de característica completa q e $\boldsymbol{\xi}$ é um vector de dimensão r previamente especificado.

Os testes mais usuais para testar as hipóteses acima referidas são o *Teste de Wald*, o *Teste de Wilks* (também denominado por Teste de Razão de Verosimilhanças) e o *Teste de Rao* (ou teste de Score).

Seguidamente ir-se-á apresentar detalhadamente os dois primeiros teste mencionados, por serem os testes mais usuais.

Teste de Wald

A *estatística de Wald*, $\mathcal{W}$, é baseada na normalidade assimptótica do estimador de máxima verosimilhança de $\boldsymbol{\beta}$.

Dado que o vector $\boldsymbol{C}\widehat{\boldsymbol{\beta}}$ é uma transformação linear de $\widehat{\boldsymbol{\beta}}$ então, pelas propriedades da distribuição Normal Multivariada e $\mathcal{I}(\boldsymbol{\beta})$ matriz de informação de Fiscer, tem-se que:

$$\boldsymbol{C}\hat{\boldsymbol{\beta}} \quad \sim \quad N_q\left(\boldsymbol{C}\boldsymbol{\beta}, \mathbf{C}\,\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})\,\mathbf{C}^T\right) \tag{3.15}$$

e, sob a hipótese nula, a estatística

$$\mathcal{W} = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi})^T[\mathbf{C}\,\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})\,\mathbf{C}^T]^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi}) \tag{3.16}$$

tem uma distribuição assimptótica χ^2, com q graus de liberdade.

Assim, ao nível de significância α, a hipótese nula é rejeitada, se o valor da estatística for superior ao quantil de probabilidade $1-\alpha$ de um χ^2_q.

Para o teste de hipóteses referido em (3.12), designando por σ_{jj} o j-ésimo elemento da diagonal de $\mathcal{I}^{-1}(\boldsymbol{\beta})$, a *Estatística de Wald*, resume-se a,

$$\mathcal{W} = (\widehat{\beta}_j - \beta_j)^T[\sigma_{jj}]^{-1}(\widehat{\beta}_j - \beta_j)$$

pelo que, sob H_0,

$$\mathcal{W} = \frac{\widehat{\beta}_j^2}{\sigma_{jj}} \quad \sim \quad \chi^2_1$$

Assim, ao nível de significância α, a hipótese nula é rejeitada, se o valor observado da estatística for superior ao quantil de probabilidade $1-\alpha$ de um χ^2_1.

Em geral, a *estatística de Wald* é a mais utilizada para testar hipóteses nulas sobre componentes individuais, ainda que também se use para testar hipóteses do tipo $\boldsymbol{\beta}_r = 0$, quando o subvector $\boldsymbol{\beta}_r$ representa o vector correspondente a uma recodificação de uma variável policotómica. Esta estatística é muito útil na comparação de modelos quando se começa a formar o modelo maximal (modelo que contém o maior número de parâmetros) e depois se consideram modelos alternativos excluíndo covariáveis, devido, essencialmente, à utilização da estimativa não restrita de máxima verosimilhança.

Teste de Razão de Verosimilhanças

A *Estatística de Razão de Verosimilhanças*, também conhecida por *Estatística de Wilks*, é definida por:

$$\Lambda = -2\ln\frac{\max_{H_0} L(\boldsymbol{\beta})}{\max_{H_0\cup H_1} L(\boldsymbol{\beta})} = -2\{l(\overline{\boldsymbol{\beta}}) - l(\widehat{\boldsymbol{\beta}})\} \tag{3.17}$$

onde $\overline{\boldsymbol{\beta}}$, o estimador de máxima verosimilhança restrito, é o valor de $\boldsymbol{\beta}$ que maximiza a verosimilhança sob a hipótese nula e $l(\cdot)$ corresponde ao máximo da função log-verosimilhança.

O Teorema de Wilks estabelece que, sob certas condições de regularidade,ver [Gey12], a estatística Λ tem, sob a hipótese nula, uma distribuição assintótica de um χ^2, onde o número de graus de liberdade é igual à diferença entre o número de parâmetros a estimar sobre $H_0 \cup H_1$ (neste caso p) e o número de parâmetros a estimar sob H_0 (neste caso $p-r$).

Assim, sob H_0,

$$\Lambda = -2\{l(\overline{\boldsymbol{\beta}}) - l(\widehat{\boldsymbol{\beta}})\} \quad \sim \quad \chi^2_q. \tag{3.18}$$

Consequentemente, ao nível de significância α, a hipótese nula é rejeitada, se o valor da estatística Λ for superior ao quantil de probabilidade $1-\alpha$ de um χ^2_q.

A *Estatística de Razão de Verosimilhanças* é a mais utilizada para comparar modelos que estão encaixados, isto é, modelos em que um é submodelo de outro.

No âmbito deste trabalho os casos particulares dos Modelos Lineares Generalizados, como a Regressão Logística e a Regressão Beta, assumem um papel preponderante na estimação do risco de crédito de um cliente. Neste sentido nas secções seguintes apresentar-se-á uma análise detalhada de cada umas destas regressões.

3.7 Modelo de Regressão Logística

A função Logística surgiu em 1789, com os estudos de crescimento populacional de Malthus. Segundo [Cra02], Alphonse Quetelet, astrónomo Belga, e o seu discípulo Pierre-François Verhust (1804-1849), 40 anos depois, recuperaram a ideia de Malthus para descrever o crescimento populacional em França, Bélgica e Rússia antes de 1833. Apesar de estar encontrada a ideia básica do modelo logístico, só em 1845, Pierre-François Verhust publicou a formulação utilizada nos estudos de crescimento da população a que chamou de função logística.

Ainda no séc. XIX, a mesma função foi utilizada para descrever as reacções químicas autocatalíticas, mas na maior parte do século esteve esquecido e só foi redescoberto em 1920 por Raymond Pearl, discípulo de Karl Peason, e Lowell Reed que o aplicaram

igualmente ao estudo do crescimento da população dos Estados Unidos da América. O primeiro estudo académico que aborda a regressão no domínio de *Credit Scoring* foi publicado em 1980 e, desde então, tornou-se a técnica estatística de eleição nos desenvolvimentos de modelos de *Credit Scoring*.

O modelo de Regressão Logística é um caso particular dos Modelos Lineares Generalizados e especialmente útil para modelar dados binários. É frequentemente utilizada em ciências médicas e sociais; no domínio dos seguros; em instituições financeiras, tendo ainda outras designações como modelo logístico, modelo logit e classificador de máxima entropia[1].

Trata-se de uma técnica estatística utilizada para produzir, a partir de um conjunto de observações, um modelo que permite a predição dos valores de uma variável categórica, frequentemente binária, a partir de um conjunto de variáveis explicativas contínuas e/ou categóricas. Nos modelos de *Credit Scoring*, a variável dependente, ocorrência de *default*, é de natureza binomial ou dicotómica, ou seja, pode apenas assumir dois valores, zero ou um, sendo que um cliente incumpridor é representado pelo valor 1.

Assim, a Regressão Logística, trata-se de um modelo de regressão para variáveis dependentes (ou resposta) binomialmente distribuídas, $Y_i \sim B(1, \pi_i)$, onde π_i é a probabilidade de sucesso para Y. É um modelo linear generalizado, $Y_i = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon_i$, onde $(x_1, ..., x_p)^T$ é um vector que corresponde às variáveis explicativas, $(\beta_1, ..., \beta_p)^T$ um vector de parâmetros e ε_i um vector de erros aleatórios. Este modelo usa como função de ligação a função *logit*:

$$\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right).$$

Podemos supor que temos n variáveis resposta independentes, ver [TS00], e $Y_i \sim B(1, \pi_i)$ ou $Y_i \sim Ber(\pi_i)$, ou seja,

$$f(y_i|\pi_i) = {\pi_i}^{y_i}(1-\pi_i)^{1-y_i} \quad , \quad y_i = 0, 1 \quad , \quad i = 1, ..., n$$

e que, a cada indivíduo i está associado um vector de covariáveis $\boldsymbol{x}_i$, $i = 1, ..., n$.

Como $\mathbb{E}[Y_i] = \pi_i$ e se tem para esta regressão $\theta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$, fazendo $\theta_i = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, conclui-se que a associação entre o valor esperado da variável resposta e as covariáveis é feita através da função de ligação canónica, função *logit*. Assim, a probabilidade de sucesso, $\pi_i = P[Y_i = 1|X = x_i]$, está relacionada com o vector $\boldsymbol{x}_i$ através de

$$\pi_i = \frac{exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1 + exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}. \tag{3.19}$$

[1] medida da desordem de um sistema

Portanto, $Logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \log(e^{\theta_i}) = \theta_i$ e

$$Logit(\pi_i) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p.$$

Como os valores possíveis de π_i se situam no intervalo $[0, 1]$, o valor de π_i é frequentemente interpretado como a probabilidade de *default*. A principal vantagem da Regressão Logística é a capacidade de estimar as probabilidades individuais de cada cliente entrar em incumprimento, sendo este um dos objectivos deste trabalho.

Odds Ratio

Odds, ou razão de chance, e probabilidade são expressões que contêm a mesma informação mas expressam-se de maneiras diferentes. A probabilidade de um acontecimento é definida através da proporção de acontecimentos favoráveis sobre o número total de acontecimentos (Lei de Laplace), enquanto que o *Odds* representam uma razão de probabilidades. Assim, sendo A um acontecimento de uma amostra aletatória, tem-se que:

$$O(A) = \frac{P(A)}{1-P(A)} \quad e \quad P(A) = \frac{O(A)}{1+O(A)}.$$

Define-se *Odds Ratio* ou quociente de razões de chances relativo a dois eventos A e B ao quociente das respectivas *Odds* e denota-se habitualmente por θ. Assim a *Odds Ratio* dos eventos A e B será definida como

$$\theta_{A,B} = \frac{O_A}{O_B} = \frac{P(A)}{1-P(A)} / \frac{P(B)}{1-P(B)} = \frac{P(A)}{P(\overline{A})} \frac{P(B)}{P(\overline{A})} = \frac{P(A)P(\overline{B})}{P(B)P(\overline{A})}.$$

O *Odds Ratio* é uma medida antiga, tendo sido usada por Snow num clássico trabalho de identificação do factor de risco de propagação da cólera em Londres, em 1853. É utilizado como medida de associação em estudos de caso-controlo.
Considerando a Regressão Logística, sendo π_i a probabilidade de sucesso de um evento, neste caso de um cliente vir a ser incumpridor, o *odds* define-se como:

$$odds_i = \frac{\pi_i}{1-\pi_i} \quad , \quad i = 1, ..., n$$

e, atendendo à definição de *odds ratio*, pode-se definir *log-odds*, à semelhança da função de ligação,

$$Logit(\pi_i) = \log(odds_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right).$$

Como um dos objectivos deste trabalho é estudar a probabilidade de *default* de um cliente, então esta medida será útil para comparar clientes com características diferentes. A *odds ratio* entre os níveis de uma covariável pode ser interpretada como o aumento estimado na probabilidade de sucesso aquando do aumento de uma unidade no valor

predito dessa mesma variável, no caso de variáveis contínuas, mantendo todas as restantes covariáveis constantes. Se a variável for categórica, a comparação é efectuada com base nos níveis da mesma, como referido em [Nun11].

3.8 Modelo de Regressão Beta

A análise de Regressão Beta, aprofundada em [FCN04], é útil para modelar variáveis contínuas que assumem valores no intervalo $]0,1[$, como ocorre, por exemplo, com taxas e proporções. A Regressão Beta é desenvolvida assumindo que a variável resposta segue uma distribuição Beta, sendo esta uma distribuição muito flexível para modelar proporções, uma vez que a sua função densidade pode tomar formas bastante distintas, dependendo dos valores dos seus parâmetros.

Nos casos em que a variável resposta assume valores no intervalo $[a,b]$ (com $a < b$ conhecidos e a, b $\in \mathbb{R}$), em [FCN04] é sugerida uma transformação da variável resposta de $Y_i = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon_i$ para $\frac{y_i - a}{b-a}$, para que se possa usar a Regressão Beta. Mas se a variável resposta assume os valores extremos 0 e 1, uma transformação útil, é $\frac{y_i \times (n-1) + 0.5}{n}$, sendo n a dimensão da amostra, também referida em [FCN04]. Esta sugestão ser-nos-á útil na formulação de alguns modelos de ajustamento realizados adiante neste trabalho.

A Regressão Beta baseia-se numa parametrização alternativa da função densidade Beta, em termos da média das variáveis e do parâmetro de precisão. Usualmente, para $Y \sim Beta(p,q)$, a função densidade Beta é expressa como:

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} \quad , \quad 0 < y < 1, \tag{3.20}$$

onde $p, q > 0$ e $\Gamma(\cdot)$ é a conhecida função Gama.

[FCN04], tendo em atenção a modelação recorrendo a modelos lineares generalizados, proposeram uma reparametrização da função de densidade descrita anteriormente, definindo $\mu = \frac{p}{p+q}$ e $\phi = p + q$, obtendo-se:

$$f(y;\mu,\phi) = \frac{\phi}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \quad , \quad 0 < y < 1, \tag{3.21}$$

com $0 < \mu < 1$ e $\phi > 0$.

Assim $Y \sim Beta(\mu,\phi)$ e tem-se que $\mathbb{E}[Y] = \mu$ e $\mathbb{V}[Y] = \frac{\mu(1-\mu)}{1+\phi}$.

O parâmetro ϕ é conhecido como parâmetro de precisão, uma vez que, para μ fixo, ϕ é maior quanto menor for a variância de$Y$ e ϕ^{-1} é designado como parâmetro de dispersão.

Considerando $Y_1, ..., Y_n$ variáveis independentes, tais que $Y_i \sim$ *Beta*(μ, ϕ), $i = 1, ..., n$, a Regressão Beta é definida como,

$$h(\mu_i) = \sum_{j=1}^{n} \boldsymbol{x}_j^T \boldsymbol{\beta}_j = \eta_i \tag{3.22}$$

onde $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ é um vector de dimensão $(k \times 1)$ de parâmetros de regressão desconhecidos $(k < n)$, $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ representa o vector de covariáveis e η_i é o predictor linear. $h(\cdot)$ é a função de ligação, estritamente monótona e diferenciável de ordem dois. Alguns exemplos de função de ligação usais na Regressão Beta são as funções *logit*: $h(\mu) = \log(\frac{\mu}{1-\mu})$, *probit*: $h(\mu) = \Phi^{-1}(\mu)$ (onde $\Phi(\cdot)$ é função de distribuição Normal), *log-log*: $h(\mu) = \log(\log(\mu))$ e *Cauchy*: $h(\mu) = tg(\pi(\mu - 0.5))$.

Neste trabalho, nos ajustamentos que se apresentam no capítulo 4, tomar-se-á como como função de ligação a função *logit*, tendo-se que:

$$\mu_i = \frac{exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \quad , \quad i = 1, ..., n \tag{3.23}$$

com as mesma definições apresentadas acima.

3.9 Qualidade de Ajustamento

Uma vez encontrado um modelo adequado de ajustamento aos dados e testada a significância dos parâmetros incluídos no modelo, surge a questão da qualidade do modelo adoptado. Quando se trabalha com muitas covariáveis, tem-se interesse em saber qual o modelo que, com menor número de variáveis explicativas, oferece uma melhor interpretação do problema em questão e que ainda se ajuste bem aos dados. O teste que indica se o modelo adoptado é o "melhor modelo" é baseado no valor da *Função Desvio*.

Dado que no processo de selecção do modelo há uma série de modelos que são tidos em consideração, comece-se por descrever os dois tipos de modelos mais usualmente referidos: o *Modelo Completo* e o *Modelo Nulo*.

O Modelo Completo ou Saturado corresponde a um Modelo Linear Generalizado com tantos parâmetros $\mu_1, ..., \mu_n$ quantas as observações $y_1, ..., y_n$. Neste modelo, os valores ajustados $\widehat{\mu}_i$ e as observações y_i confudem-se entre si pois o modelo ajusta-se perfeitamente aos dados, ou seja, as estimativas de máxima verosimilhança dos μ_i são as próprias observações, isto é, $\widehat{\mu}_i = y_i$.

O Modelo Nulo é o modelo mais simples, em que se considera apenas um parâmetro, que representa a média μ, comum a todas as observações y_i. É um modelo simples mas que raramente captura a estrutura inerente aos dados.

3.9.1 Função Desvio

Na prática, pretende-se encontrar um modelo cujo número de parâmetros se encontre entre o número de parâmetros de cada um dos modelos acima descritos. A Função Desvio irá auxiliar na escolha do modelo a adoptar.

Sejam $\widehat{\boldsymbol{\beta}}_E$ e $\widehat{\boldsymbol{\beta}}_S$ as estimativas de máxima verosimilhança para o modelo em estudo e o modelo saturado, respectivamente.

O quociente

$$\lambda = \frac{l(\widehat{\boldsymbol{\beta}}_E)}{l(\widehat{\boldsymbol{\beta}}_S)}$$

mede o afastamento entre as verosimilhanças dos modelo acima referidos, pelo que quanto menor o valor de λ, melhor será o modelo ajustado e um valor muito elevado indicará um ajustamento de fraca qualidade.

Logaritmizando a expressão anterior, obtém-se

$$\log \lambda = l(\widehat{\boldsymbol{\beta}}_E) - l(\widehat{\boldsymbol{\beta}}_S) \tag{3.24}$$

com $l(\widehat{\boldsymbol{\beta}}_S)$ e $l(\widehat{\boldsymbol{\beta}}_E)$ o máximo da função log-verosimilhança para o modelo saturado e o modelo em estudo, respectivamente.

O **desvio** é, portanto, definido como uma medida de afastamento entre o modelo saturado e o modelo ajustado, calculado através da expressão:

$$D^* = 2\left[l(\widehat{\boldsymbol{\beta}}_S) - l(\widehat{\boldsymbol{\beta}}_E)\right].$$

Mostra-se, ver [Dob02], que

$$D^* = 2\left[l(\widehat{\boldsymbol{\beta}}_S) - l(\widehat{\boldsymbol{\beta}}_E)\right] \quad \sim \quad \chi^2_{n-p}.$$

Considere-se, agora, a função log-verosimilhança de um Modelo Linear Generalizado, ver (3.2) com a finalidade de se especificar a Função de Desvio e o Desvio Reduzido,

$$\log L(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{\phi}(y_i q(\mu_i) - b(q(\mu_i))) + c(y_i, \phi) \tag{3.25}$$

em que $q(\mu_i)$ representa a relação funcional entre θ_i e μ_i.

Como para o modelo saturado se tem $\widehat{\mu}_i = y_i$ e sendo $\widehat{\mu}$ a estimativa de máxima verosimilhança de μ_i para o modelo em estudo, então o **Desvio Reduzido** é obtido através de:

$$\begin{aligned} D^*(\mathbf{y}, \boldsymbol{\mu}) &= -2\left(l(\widehat{\boldsymbol{\beta}}_M) - l(\widehat{\boldsymbol{\beta}}_S)\right) \\ &= -2\sum_{i=1}^{n} \frac{1}{\phi}\left([y_i q(\widehat{\mu}_i) - b(q(\widehat{\mu}_i))] - [y_i q(y_i) - b(q(y_i))]\right) \\ &= \frac{D(\mathbf{y}, \boldsymbol{\mu})}{\phi} \end{aligned}$$

sendo $D(\mathbf{y}, \boldsymbol{\mu})$ o **Desvio** para o modelo em análise, dado por:

$$D^*(\mathbf{y}, \boldsymbol{\mu}) = -2\sum_{i=1}^{n}\left([y_i q(\widehat{\mu}_i) - b(q(\widehat{\mu}_i))] - [y_i q(y_i) - b(q(y_i))]\right). \tag{3.26}$$

É de notar que a Função Desvio pode ser interpretada como soma ponderada das distâncias entre as estimativas para os valores médios $\widehat{\mu}_i$ e as observações y_i, sendo ainda possível decompor a função desvio como

$$D^*(\mathbf{y}, \boldsymbol{\mu}) = \sum_{i=1}^{n} d_i \tag{3.27}$$

ou seja, pode ser decomposta como a soma de parcelas d_i que medem a diferença dos logaritmos das verosimilhanças observada e ajustada para cada observação. A soma destas componentes é assim uma medida da discrepância total entre as duas log-verosimilhanças.

Verifica-se, facilmente, que o Desvio é uma função não negativa, sendo que para o modelo saturado toma o valor zero, e vai crescendo à medida que as covariáveis vão sendo retiradas do modelo.

Uma outra propriedade do Desvio é a aditividade para modelos encaixados. Suponha-se que M_1 e M_2 são dois modelos intermédios, com M_2 encaixado em M_1, ou seja, são modelos do mesmo tipo, mas o modelo M_2 contém menos parâmetros que o modelo M_1. Designando $D(y; \widehat{\boldsymbol{\mu}}_j)$ o desvio do modelo M_j, $j = 1, 2$, então a estatística da razão de verosimilhanças para comparar estes dois modelos resume-se a

$$-2(l_{M_2}(\boldsymbol{\beta}_2) - l_{M_1}(\boldsymbol{\beta}_1)) = \tfrac{D(\mathbf{y};\widehat{\boldsymbol{\mu}}_2) - D(\mathbf{y};\widehat{\boldsymbol{\mu}}_1)}{\phi}.$$

Sob a hipótese do modelo M_1 ser verdadeiro, tem-se

$$\tfrac{D(\mathbf{y};\widehat{\boldsymbol{\mu}}_2) - D(\mathbf{y};\widehat{\boldsymbol{\mu}}_1)}{\phi} \quad \sim \quad \chi^2_{p_1 - p_2},$$

onde p_j representa a dimensão do vector $\boldsymbol{\beta}$ para o modelo M_j, $j = 1, 2$. A comparação

de modelos encaixados pode assim ser feita com base da diferença dos desvios de cada modelo.

Para os dois casos particulares dos Modelos Lineares Generalizados descritos neste capítulo apresentam-se em seguida os resultados da *Função Desvio* para ambas as regressões.

No caso da Regressão Logística, tem-se que:

$$D(\mathbf{y}, \widehat{\boldsymbol{\pi}}) = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\widehat{\pi}_i}\right) + (1-yi)\log\left(\frac{1-y_i}{1-\widehat{\pi}_i}\right)\right] \tag{3.28}$$

com $Y_i = \beta_0 + \beta_1 x_1 + ... + \beta_i x_i + \varepsilon_i$ e $\widehat{\pi}_i = \frac{exp(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}})}{1+exp(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}})}$.

No caso da Regressão Beta,

$$D(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\phi}) = \sum_{i=1}^{n} 2(l(\widetilde{\mu}_i, \phi) - l(\mu_i, \phi)) \tag{3.29}$$

onde $l(\cdot)$ é a função de máxima verosimilhança para o modelo e $\widetilde{\mu}_i$ é o resultado de $\frac{\partial l(\mu_i,\phi)}{\partial \mu_i} = 0$. Se ϕ for um parâmetro conhecido, a função desvio é $D(\mathbf{y}, \overline{\mu}, \phi)$, onde $\overline{\mu}$ é estimador da função de máxima verosimilhança.

3.10 Análise de Resíduos

A análise de resíduos consiste num conjunto de técnicas utilizadas para aferir a adequabilidade de um modelo aos dados, são técnicas destinadas a verificar a validade das hipóteses efectuadas sobre o modelo, nomeadamente no que diz respeito à escolha da distribuição, da função de ligação e de termos do preditor linear, como também para para ajudar a verificar se há observações mal ajustadas, isto é, que não são bem explicadas pelo modelo. Um resíduo R_i deve exprimir a discrepância entre o valor observado y_i e o valor $\widehat{\mu}_i$ ajustado pelo modelo.

A escolha mais comum para avaliação dos resíduos corresponde aos *Resíduos de Pearson*, definidos por:

$$R_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\mathbb{V}[Y_i]}}}. \tag{3.30}$$

Estes resíduos apresentam, no entanto, a desvantagem de ter uma distribuição bastante assimétrica para modelos não normais.

Os *Resíduos Standartizados de Pearson* introduzem uma correcção aos resíduos de Pearson, para fazer face ao facto de $\widehat{\mu}_i$ serem parâmetros estimados, e são calculados como

$$R^{*P}_i = \frac{R^P_i}{\sqrt{\phi(1-h_{ii})}} \tag{3.31}$$

sendo h_{ii} o elemento da diagonal principal da matriz Hessiana

$$H = \mathbf{D}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{D}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}^{1/2} \tag{3.32}$$

onde $\mathbf{D} = diag(\frac{1}{\phi}\mathbb{V}[\mu_i])$.

Pode ainda considerar-se um outro tipo de resíduo, baseado na função desvio, denominado por *Desvio Residual*:

$$R^{*D}_i = \frac{R^D_i}{\sqrt{\hat{\phi}(1-h_{ii})}}, \tag{3.33}$$

com $R^D_i = \delta_i\sqrt{d_i}$, em que $\delta_i = sinal(y_i - \hat{\mu}_i)$ e d_i são elementos da função de desvio.

A análise de resíduos, nomeadamente no que diz respeito a adaptação de um modelo, pode ser realizada através de uma avaliação informal dos resíduos, ou seja, analisando o gráfico dos valores dos resíduos contra os valores ajustados, o que nos permite analisar mais facilmente a existência de *outliers* que poderão não ser incluídos no modelo.

3.11 Selecção dos Modelos

Quando se ajusta um modelo a uma variável resposta e se pretende encontrar o melhor modelo, por vezes passa-se pelo processo de adição ou remoção de covariáveis ao modelo inicialmente considerado. Os Modelos Lineares Generalizados contêm diversos fenómenos aleatórios modelados, geralmente, a partir um número elevado de covariáveis que podem ser potencialmente importantes para explicar a variabilidade entre os dados. Também tem interesse estudar a influência de possíveis iterações entre as covariáveis.

Por vezes, adicionar uma variável explicativa ao modelo incorpora mais informação, o que se traduz num melhor ajustamento. No entanto, estatísticamente, verifica-se que cada variável não necessária ao modelo (que não incorpora informação útil ao modelo), diminui a precisão das estimativas para os parâmetros de regressão.

A inclusão de variáveis explicativas no modelo aumenta o número de parâmetros p a estimar e diminui os erros $y_i - \widehat{y}_i$ do ajustamento. Um valor elevado de p tende a aumentar a qualidade do ajustamento, diminuindo os desvios, embora quantos mais parâmetros forem necessários estimar, maior será a variância das estimativas de β_j, $j = 1, ..., p$.

Os critérios, medidas e metodologias que se seguem pretendem dar algumas indicações acerca da inclusão/exclusão de variáveis, até que se decida pelo modelo mais adequado.

Nesta secção segue-se de perto a dissertação [Val10].

3.11.1 Método Stepwise para a escolha das covariáveis

O método *stepwise* é utilizado, principalmente, quando se quer considerar, no início, um número relativamente elevado de variáveis independentes para incluir na função, mas também pode ser em modelos iniciais nulos. A cada passo as variáveis menos úteis - que menos explicitam o modelo - são eliminadas, ou no caso de se iniciar com o modelo nulo, as variáveis mais significativas são adicionadas, e apenas são retiradas as covariáveis que menos explicam o modelo.

Os métodos *stepwise* podem ter duas dinâmicas diferentes; a *forward stepwise* e b*ackward stepwise*. Resumidamente, o método *backward stepwise* parte de um modelo inicial com todas as possíveis variáveis, que vão sendo eliminadas a cada passo até ser atigindo o modelo final. O método *forward stepwise* inicia-se com um modelo sem nenhuma variável explicativa (modelo nulo) e a cada passo são incluídas as variáveis relevantes até a obtenção do modelo final.

O método *stepwise* baseia-se no valor dos *p-values* relativos aos testes de razão de verosimilhanças de Wilks, entre modelos com inclusão ou exclusão de variáveis explicativas, para definir quais as covariáveis que devem ser incluídas do modelo final. Em suma, o método processa-se da seguinte forma: começa por se calcular o valor do *p-value* obtido pelo teste de Wald e, com base neste, escolhe-se qual a variável que deve sair (ou entrar) no modelo final. Quanto menor (ou maior) for o valor do *p-value* mais (menos) importante é considerada a covariável. Após a escolha da covariável, faz-se uma segunda análise ao seu grau de importância através do valor do *p-value* do teste de razão de verosimilhanças entre os modelos que a incluem e excluem, e assim se toma uma decisão acerca da exclusão (ou inclusão) da variável no modelo final.

3.11.2 Critério de Informação Akaike

Existem outros critérios que podem ser utilizados para a selecção de modelos, critérios que têm em consideração a complexidade do modelo. Estes baseiam-se, essencialmente, na penalização na função de log-verosimilhança, introduzindo um factor de correcção como modo de penalização da complexidade do modelo.

O Critério de Informação de Akaike foi desenvolvido por Hirotugu Akaike, sob o nome de "*Akaike Information Criterion*" (AIC), em 1971, e foi proposto por Akaike em 1974, sendo uma medida de equilíbrio entre a qualidade do ajustamento e o número de parâmetros incluídos no modelo. É fundamentado no conceito de entropia, e oferece uma medida relativa à informação perdida quando um determinado modelo é utilizado. Pode ser usado para descrever o equílibrio entre a variância e a tendenciosidade (*bias*) da construção do modelo ou, por outras palavras, entre a precisão e a complexidade do modelo.

O AIC não é um teste ao modelo no sentido de testar hipóteses, mas sim um teste entre modelos, ou seja, é uma ferramenta para seleccionar um modelo de entre um conjunto de modelos. O AIC classifica os modelos e o que tiver o menor AIC deve ser considerado o melhor modelo.

A estatística correspondente para o modelo em H_0 é,

$$AIC = -2l(\boldsymbol{\beta}; \phi, \mathbf{y}) + 2p \tag{3.34}$$

onde $l(\boldsymbol{\beta}; \phi, \mathbf{y})$ corresponde à função de log-verosimilhança que se tem vindo a considerar.

Em suma, um ajustamento de boa qualidade traduz-se num valor elevado para a função verosimilhança, com o menor número de parâmetros possível, logo pode concluir-se que o ajustamento é tanto melhor quanto menor for o valor de AIC.

3.12 Observações discordantes

A análise dos resíduos permite averiguar a existência de desvios sistemáticos do modelo. No entanto, também é interessante investigar se existem desvios isolados do modelo, isto é, observações mal ajustadas, que se distinguem das outras por não seguirem o mesmo padrão, sendo tais observações denominadas por *observações discordantes*. Na análise dos desvios isolados existem três noções importantes: a *repercussão*, a *influência* e a *consistência*.

Medida de repercussão ("leverage")

A medida de repercussão mede o efeito que a observação tem nos valores preditos. Assim pode-se considerar que um ponto tem repercussão elevada se

$$h_{ii} > \frac{2p}{n}, \tag{3.35}$$

onde $p = \sum_{i=1}^{n} h_{ii}$ e $h_{ii} \in \mathbf{H}$, ver (3.32).

Medida de influência

Uma observação diz-se influente se uma sua ligeira modificação ou a sua exclusão do modelo produz alterações significativas no ajustamento do modelo. Uma medida frequente para a influência de uma observação i é a distância de Cook. A distância de Cook representa o efeito de excluir uma dada observação e, em valor absoluto, é dada por:

$$C_i = |R*_i{}^P|(\frac{n-p}{p} \cdot \frac{h_i}{1-h_i})^{1/2}, \tag{3.36}$$

em que R^{*P}_i é o resíduo indicado em (3.31), $p = \sum_{i=1}^{n} h_{ii}$ e $h_{ii} \in \mathbf{H}$ (ver (3.32)).

Medida de consistência

Uma observação inconsistente é, em geral, uma observação com um resíduo elevado. Esta inconsistência pode ser devida a um valor extremo da variável resposta ou de uma ou mais covariáveis. Uma observação consistente deve seguir a tendência sugerida pelas restantes observações, no entanto, pode haver observações consistentes com repercussões elevadas. Para se estudar a existência de observações inconsistentes utiliza-se um tipo de resíduo, o *Resíduo de Verosimilhança*, dado por:

$$R_i{}^G = \delta_i \sqrt{(1-h_{ii})(R^{*D}_i)^2 + h_{ii}(R^{*P}_i)^2}, \tag{3.37}$$

onde $\delta_i = sinal(y_i - \hat{\mu}_i)$, $h_{ii} \in \mathbf{H}$ (ver (3.32)) e R^{*P}_i e R^{*D}_i os resíduos standartizados de Pearson e os devios residuais, referidos em (3.31) e (3.33), respectivamente.

Assim, observações com valores elevados de R_{Gi} podem ser consideradas inconsistentes. Gráficos de R^G_i contra i, h_{ii} ou $\hat{\eta}_i$ podem ser úteis para estudar as observações quanto à sua consistência.

Os Modelos Lineares Generalizados têm como objectivo principal estudar a relação entre variáveis, avaliar o efeito de uma ou mais variáveis explicativas ou independentes sobre uma variável dependente ou resposta, pelo que a sua metodologia é de extrema

importância neste presente trabalho.

Nesta dissertação pretende-se modelar a probabilidade de default de um cliente em função de covariáveis que caracterizam o cliente e o crédito contratado e estimar a taxa de recuperação em função das mesmas variáveis, assim ir-se-á utilizar os casos particulares dos Modelos Lineares Generalizados, a Regressão Logística e a Regressão Beta.

4

Modelação de Risco de Crédito - Aplicação

Neste capítulo pretende-se estimar o *spread* adequado ao perfil de risco de um novo cliente, aquando do momento de concessão. Para tal, recorrer-se-á ao modelo de estimação do *spread* proposto no capítulo 2 e em [EGFS14]. Tornando-se necessário efectuar a estimação da probabilidade de *default* e da taxa de recuperação para as quais se utilizam os modelos de Regressão Logística e Beta, respectivamente. No que se refere à probabilidade de *default* utilizou-se como *proxy* da variável aleatória correspondente dois possíveis critérios de classificação de ocorrência de *default*. O que permitiu uma análise do impacto da definição de cliente incumpridor no risco/*spread* associado a cada cliente. Por fim, e como análise complementar da carteira, estudou-se a proporção das prestações pagas, estimando-se, em caso de ocorrência de *default*, a percentagem do empréstimo que se encontrará pago.

4.1 A Carteira de Crédito

4.1.1 Base de dados

Os dados utilizados no estudo foram cedidos por uma instituição bancária de Cabo Verde e representam a carteira de crédito ao consumo no período compreendido entre Janeiro de 2003 e Outubro de 2011.

A base de dados contém a informação de 22.044 processos encerrados, tendo-se observado dois sub-conjuntos de características: as características / variáveis sócio-demográficas, que caracterizam o cliente no momento do empréstimo, e as variáveis sócio-económicas, que identificam as características do empréstimo.

A Tabela 4.1 ilustra as características dos clientes utiliadas como variáveis explicativas para o desenvolvimento do estudo pretendido.

Variável	**Código da Variável**	**Descrição**
Variáveis Sócio-Demográficas		
Idade	Idade	Idade do Cliente
Género	Genero	Género do Cliente
Estado Civil	Civil	Estado Civil do Cliente
Habilitações	Habilitacoes	Habilitações Literárias do Cliente
Actividade Profissional	ActProfissional	Actividade Profissional do Cliente
Entidade Patronal	EntPatronal	Tipo de Entidade Patronal do Cliente
Agência	Agencia	Localização da Agência da Instituição Bancária
Variáveis Económicas do Empréstimo		
Valor do Empréstimo	VEmprest	Montante de crédito cedido pela instituição bancária
Prazo	Prazo	N° de prestações mensais
Taxa Nominal	TxN	Taxa de juro nominal
Valor da Prestação	VPrest	Montante que o Cliente paga em cada prestação
Prestações Pagas	PrestPagas	N° de prestações liquidadas pelo Cliente no final do contrato
Tipo de Garantia	Garantia	Garantia apresentada pelo Cliente

Tabela 4.1: Definição das Variáveis

As variáveis fornecidas pela Instituição Bancária vinham categorizadas de acordo com os critérios definidos inteiramente pela instituição. Na Tabela 4.2 descreve-se as categorias de cada variável e o número de processos existentes em cada categoria.

Sabe-se que para preparação da base de dados, foram consideradas algumas restições de acordo com a Instituição Bancária, com objectivo de eliminar possíveis erros ou mesmo valores atípicos. Pelo que, se rejeitaram clientes de acordo com os seguintes critérios: montante de empréstimo inferior ou igual a 3.000 e superior a 2.500.000 ECV (Escudos Cabo-verdianos); idade inferior a 17 e superior a 70 anos; taxa nominal inferior a $2,5\%$, uma vez que são os funcionários da instituição que usufruem desta mesma taxa e superior ou igual a 40%.

Variável	Categoria	Grupo	N° de processos
Idade	1	Inferior a 27	2642
	2	Entre 27 e 28	1681
	3	Entre 29 e 32	3199
	4	Entre 33 e 42	7241
	5	Superior a 42	7281
Género	F	Feminino	9036
	M	Masculino	13008
Estado Civil	1	Casado, Divorciado, Separado	5751
	2	Solteiro, União de Facto, Viúvo, Missing	16293
Habilitações	1	Habilitações Desconhecidas	4299
	2	Escolaridade Obrigatória	11454
	3	Ensino Secundário	3999
	4	Curso Médio e formação profissional	1134
	5	Curso Superior	1158
Act.Profissional	1	Act. Desconhecida, Doméstica, Estudante e Pequena/Média Empresa	2335
	2	Outros	6762
	3	Liberal/Quadro Superior, Operário especializado ou não especializado e Quadro Médio	1536
	4	Emp. Escritório, Comércio e Serviços	11411
Entidade Patronal	1	Instituições Financeiras, Instituos e Serviços Autónomos	1639
	2	Aposentado/Pensionista, Câmara Municipal, Ministérios	11239
	3	Grandes Empresas, Hotelaria e Restauração	2899
	4	Não Declarou	3672
	5	PME, Conta Própria, Outras	2595
Agência	1	4, 7, 14, 18, 28 e 29	2176
	2	5 e 10	2646
	3	6, 23, 26 e 30	1101
	4	1, 2, 8, 9, 11 e 24	10237
	5	3, 12, 19, 22, 25, 27, 31 e 32	5884
Valor do Empéstimo	1	$\leq$ 107.600	4408
(ECV)	2	]107.600, 200.000[	4504
	3	[200.000, 512.320[	8761
	4	$\geq$ 512.320	4371
Taxa Nominal	1	< 12,5	2974
	2	$\geq$ 12,5	19070
Valor da Prestação	1	$\leq$ 6.120	5323
(ECV)	2	]6.120, 9.742]	5720
	3	]9.742, 14.570]	4472
	4	> 14.570	6529
Prestações Pagas	1	Inferior a 18	5210
	2	Entre 18 e 23	4017
	3	Entre 24 e 34	6135
	4	Superior a 34	6682
Prazo	1	Inferior a 13	4319
	2	Entre 13 e 24	3871
	3	Entre 25 e 36	6948
	4	Entre 37 e 48	4463
	5	Superior a 48	2243
Tipo de Garantia	1	Depósitos, Hipoteca s/ Imóveis, Junto da Instituição, Outras Entidades, Outras Hipotecas	7940
	2	Outras Cauções, Penhor	14104

Tabela 4.2: Descrição das Categorias

4.1.2 Definição de Cliente incumpridor

Uma das variáveis de interesse neste modelo será a qualidade de crédito de cada cliente, ou seja, a medida com probabilidade de um cliente entrar em *default*, isto é, em incumprimento. Ne extensa literatura sobre o risco de crédito, existem inúmeras definições para a váriavel resposta. Com o objectivo de analisar a probabilidade de *default* de um cliente, neste trabalho, seleccionaram-se dois critérios para definir a variável resposta.

O primeiro critério considerado, referido em [Sid06], tem por base o número de dias em atraso no pagamento das prestações do empréstimo: considera-se incumpridor o cliente que esteve pelo menos uma vez, durante o contrato de empréstimo, com mais de noventa dias de atraso no pagamento de uma prestação. O segundo critério define como cliente incumpridor aquele que, no vencimento do processo de empréstimo, tem um montante em dívida estritamente positivo, ou seja, se possui valores em dívida à data que o processo deveria ser encerrado.

Na Tabela 4.3 é apresentado um resumo de ambas as definições.

Critério	**Definição**
1° Critério	Não foram realizados pagamentos por um período superior a 90 dias, pelo menos uma vez durante o contrato
2° Critério	No vencimento do processo de empréstimo, montante em dívida estritamente positivo

Tabela 4.3: Definição de Cliente incumpridor

A Tabela 4.4 resume a classificação dos clientes da base de dados, de acordo com cada um dos critérios de incumprimento adoptados.

Critério	**Situação**	***N°***	%
1° Critério	Cumpridor	19.657	89,17%
	Incumpridor	2.387	10,83%
2° Critério	Cumpridor	17.344	78,68%
	Incumpridor	4.700	21,32%

Tabela 4.4: Definição de Cliente incumpridor

Como se pode observar na Tabela 4.3, o primeiro critério de incumpridor envolve 2.387 processos encerrados, o que representa $10,83\%$ da carteira e o segundo critério abrange 4.700 processos, o que afigura $21,32\%$ da exposição.

Desde já se pode notar que o critério adoptado para classificação de *default* é preponderante na quantificação do risco, o que se verá com mais detalhe nas secções que se seguem.

4.1.3 Análise Estatística das Variáveis

A Estatística Descritiva tem como objectivo descrever e analisar a informação que nos é fornecida, caracterizando assim o conjunto de dados de que se dispõe. Esta secção iniciar-se-á com uma análise preliminar das estatísticas descritivas e posteriormente uma análise gráfica que, muitas vezes, expõe a informação acerca das variáveis que constituem a carteira de crédito e das relações entre as mesmas, de forma mais visível.

Na Tabela 4.5, são apresentadas as frequências relativas para cada categoria das variáveis definidas anteriormente na Tabela 4.2. Uma outra forma de analisar os dados da Tabela 4.5 é recorrendo a histogramas e a diagramas de "caixa-e-bigodes".

Idade	**Género**	**Est.Civil**	**Habilitações**	**Act.Profissional**	**Ent.Patronal**	**Agência**
1: 11,99%	F: 40,99%	1: 26,09%	1: 19,50%	1: 10,59%	1: 7,44%	1: 9,87%
2: 7,63%	M: 59,01%	2: 73,91%	2: 51,96%	2: 30,68%	2: 50,98%	2: 12,00%
3: 14,51%			3: 18,14%	3: 6,97%	3: 13,15%	3: 4,99%
4: 32,85%			4: 5,14%	4: 51,76%	4: 16,66%	4: 46,44%
5: 33,03%			5: 5,25%		5: 11,77%	5: 26,69%

V.Empréstimo	**Prazo**	**Tx.Nominal**	**V.Prestação**	**P.Pagas**	**T.Garantia**
1: 10,59%	1: 19,59%	1: 13,49%	1: 24,15%	1: 26,63%	1: 36,02%
2: 30,68%	2: 17,56%	2: 86,51%	2: 25,95%	2: 18,22%	2: 63,98%
3: 6,97%	3: 31,52%		3: 20,29%	3: 27,83%	
4: 51,76%	4: 21,15%		4: 29,62%	4: 30,31%	
	5: 10,18%				

Tabela 4.5: Análise Preliminar das Variáveis

Os histogramas da Figura 4.1 ilustram os dados das variáveis sócio-demográficas.

Construiram-se ainda alguns gráficos do tipo "caixa-e-bigodes" para ilustrar algumas das variáveis quantitativas envolvidas no estudo.

A "caixa" é delimitada superior e inferiormente por dois traços que localizam a altura correspondente respectivamente ao primeiro e terceiro quartil. O traço intermédio que divide a caixa em duas partes corresponde à mediana e permite identificar assimetrias nos dados. Os dois segmentos de recta que unem verticalmente os limites superior e inferior da caixa, "os bigodes", correspondem aos valores máximo e mínimo dos dados e permitem identificar a existência de *outliers*, no caso de os bigodes serem, relativamente à caixa, muito grandes. Os outliers são identificados da seguinte forma ([MSP02]):

1. x_i é um *outlier* severo se:

$$x_i < Q_1 - 3\,(Q_3 - Q_1) \quad ou \quad x_i > Q_3 + 3(Q_3 - Q_1)$$

2. x_i é um *outlier* moderado se:

$$Q_1 - 3(Q_3 - Q_1) < x_i < Q_1 - 1,5(Q_3 - Q_1) ou Q_3 + 1,5(Q_3 - Q_1) < x_i < Q_3 + 3(Q_3 - Q_1)$$

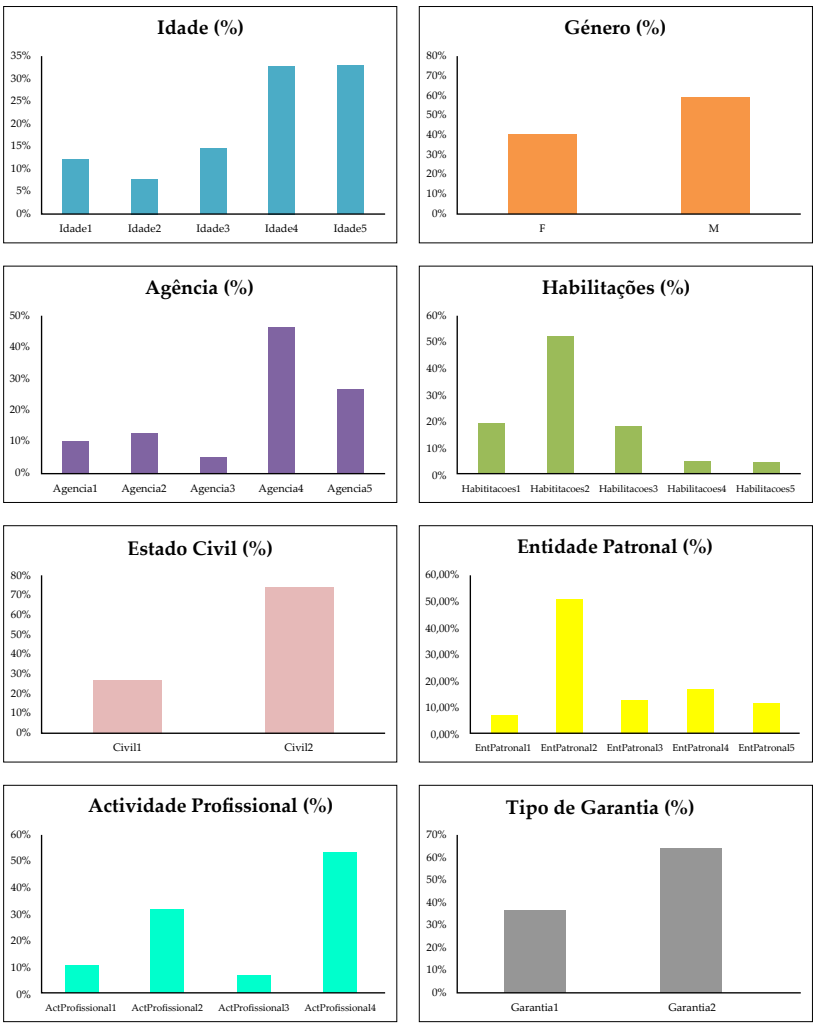


Figura 4.1: Histogramas Variáveis Qualitativas

A Figura 4.2 corresponde à representação gráfica da variável *Valor Empréstimo* e esta figura corresponde à distribuição dos valores variável. Como se pode observar existe uma notória assimetria entre os montantes de crédito, tendo em conta que a semi-caixa superior é maior. E verifica-se a existência de *outliers* severos e moderados, uma vez que o tamanho do bigode superior é superior a uma vez e meia a distância entre os quartis e superior trê vezes a mesma distância, respectivamente.

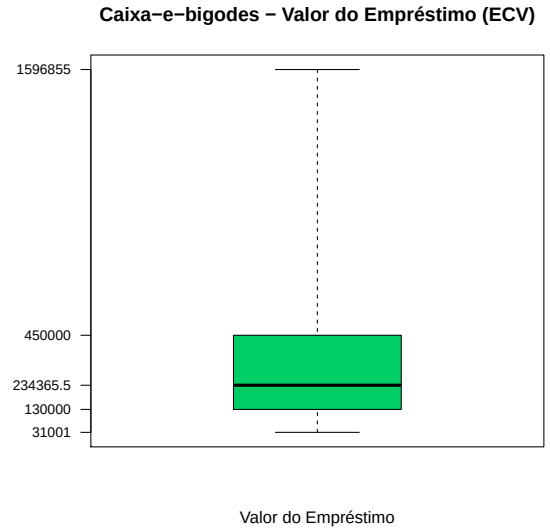
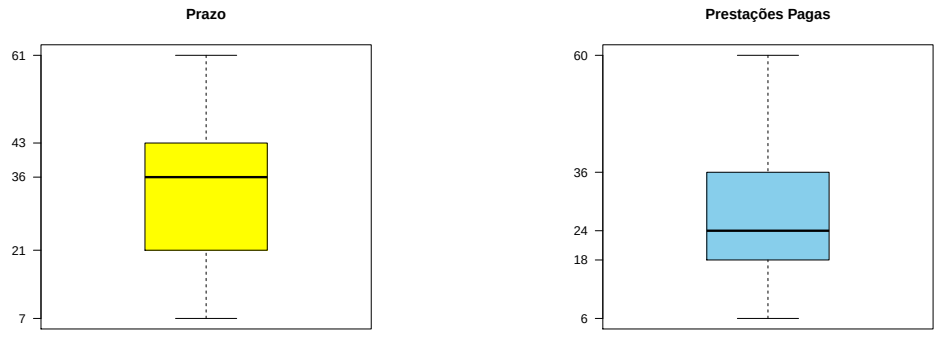


Figura 4.2: Caixa-e-Bigodes: Variável *Valor Empréstimo*

A Figura 4.3 representa o gráfico da "caixa-e-bigodes" para as variáveis *Prazo* e *Prestações Pagas* e observa-se para ambas as variáveis, assimetria dos dados. Em relação à variável *Prestações Pagas* ainda se pode concluir a existência de *outliers* moderados.

Note-se que os gráficos da Figura 4.3 são poucos ilustrativos do problema em estudo, sendo que o número de prestações pagas é, naturalmente, função do número de prestações totais contratadas no empréstimo.



(a) Prazo (b) PPagas

Figura 4.3: Caixa-e-Bigodes: *Prazo* e *Prestações Pagas*

Os histogramas da Figura 4.4 representam a frequência relativa das variáveis *Valor Empréstimo*, *Prazo*, *Prestações Pagas* e *Valor da Prestação*. Donde se pode concluir que a maioria dos emprétimos são de valor elevado, em relação às restantes variáveis a distribuição da exposição é uniforme.

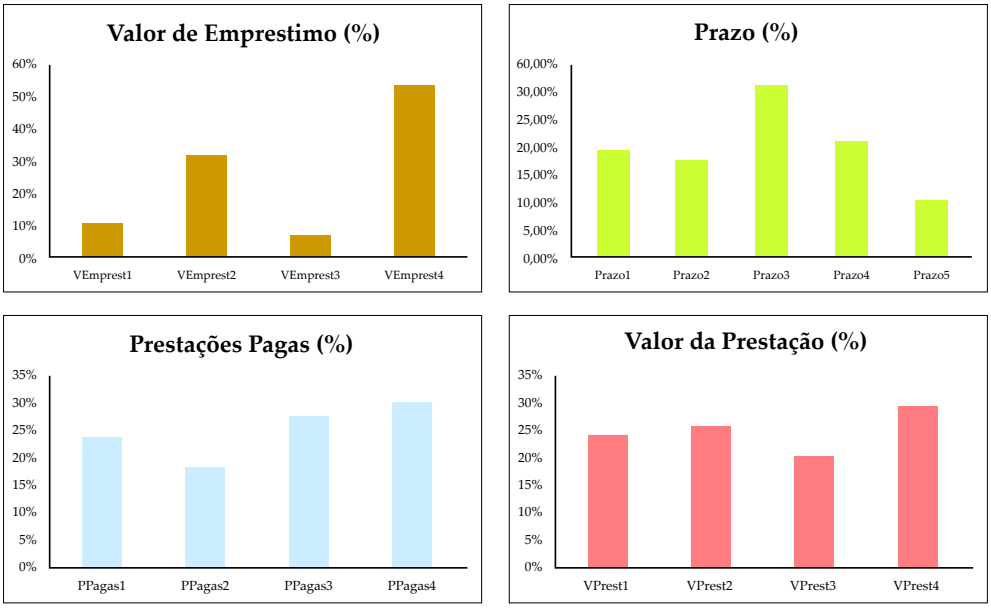


Figura 4.4: Histogramas Variáveis Quantitativas

Para ilustrar a relação entre as covariáveis e a variável resposta, a ocorrência de *default*, realizaram-se alguns gráficos do tipo "caixa-e-bigodes" com os dados da carteira de crédito. Uma outra análise realizada foi a relação entre a variável resposta *default* e as variáveis quantitativas *Valor de Empréstimo*, *Prestações Pagas* e *Prazo*. Como se adoptou duas definições de incumpridor, apresenta-se seguidamente, nas Figuras 4.5 e 4.6, os gráficos referentes ao primeiro critério definido.

Observando as Figuras 4.5 e 4.6, verifica-se através do gráfico "*Default/Valor de Empréstimo*" que os clientes que entram em *default*, são aqueles cujos montantes de crédito são mais elevados e existe um número elevado de clientes com montantes de crédito muito elevado que entram em incumprimento, observando-se um pico para valores superiores a 1.500.000 ECV. Observando agora os gráficos "*Default/Prazo*" das mesmas figuras, verifica-se que quanto maior o prazo do empréstimo, maior a probabilidade de ocorrer *default*. Para o número de *Prestações Pagas*, destaca-se a existência de *outliers* e observa-se que a ocorrência de *default* ocorre mais frequentemente quando o número de prestações pagas é elevado.

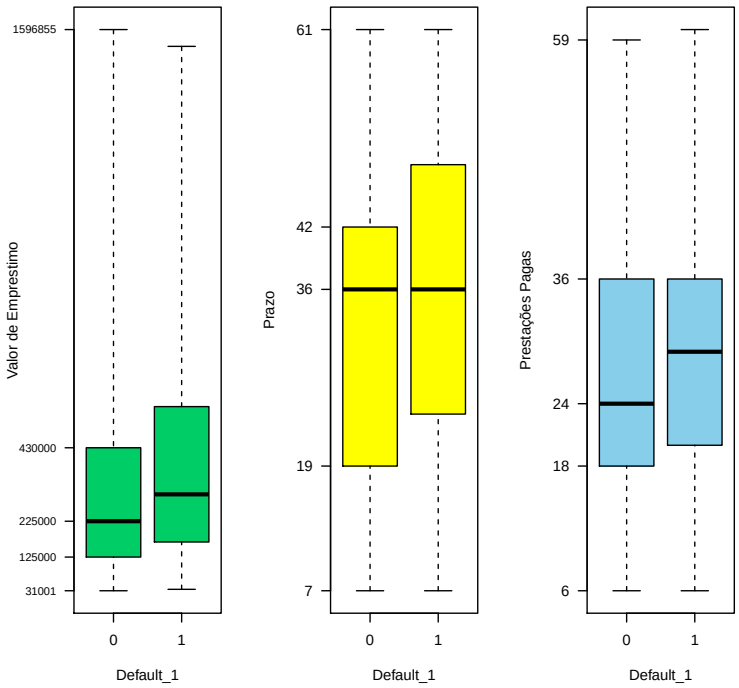


Figura 4.5: 1° Critério: Variável *default vs* variáveis Quantitativas

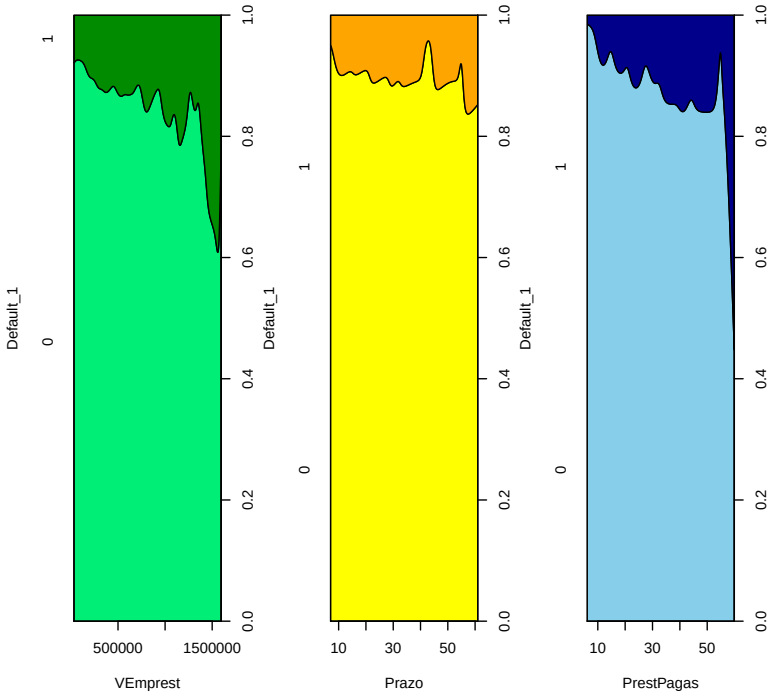


Figura 4.6: 1° Critério: Relação entre a variável *default* e as variáveis Quantitativas

De forma análoga, considerando a segunda definição de incumpridor, obtiveram-se os gráficos das Figuras 4.7 e 4.8.

Tendo em conta o 2° critério adoptado, observando as Figuras 4.7 e 4.8, verifica-se que no gráfico "*Default/Valor de Empréstimo*", os clientes que entram em *default*, são aqueles cujo montante de empreéstimo é mais baixo. Observando agora os gráficos "*Default/-Prazo*"das mesmas figuras, nota-se que a probabilidade de ocorrer *defaut* é maior quando o prazo é inferior. Em relação ao número de *Prestações Pagas*, não se observam diferenças entre os clientes que são ou não incumpridores. Como referido anteriormente, esta variável em termos absolutos é pouco descritiva do risco em estudo.

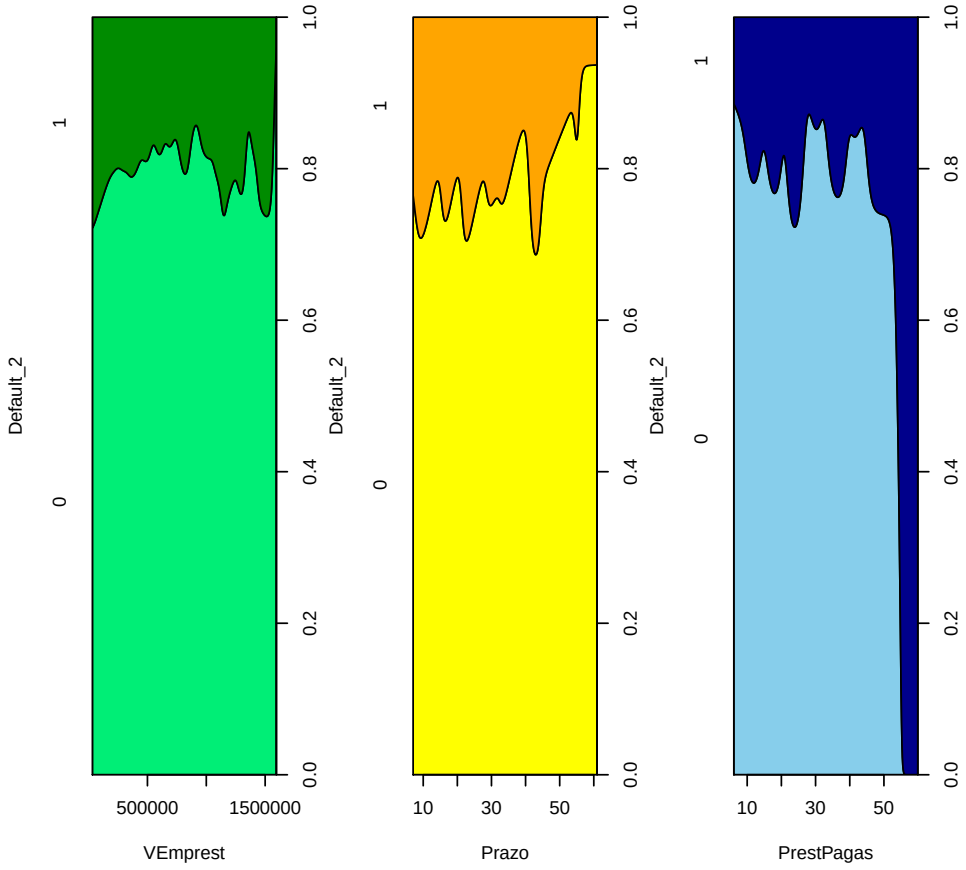


Figura 4.7: 2° Critério: Relação entre a variável *default* e as variáveis Quantitativas

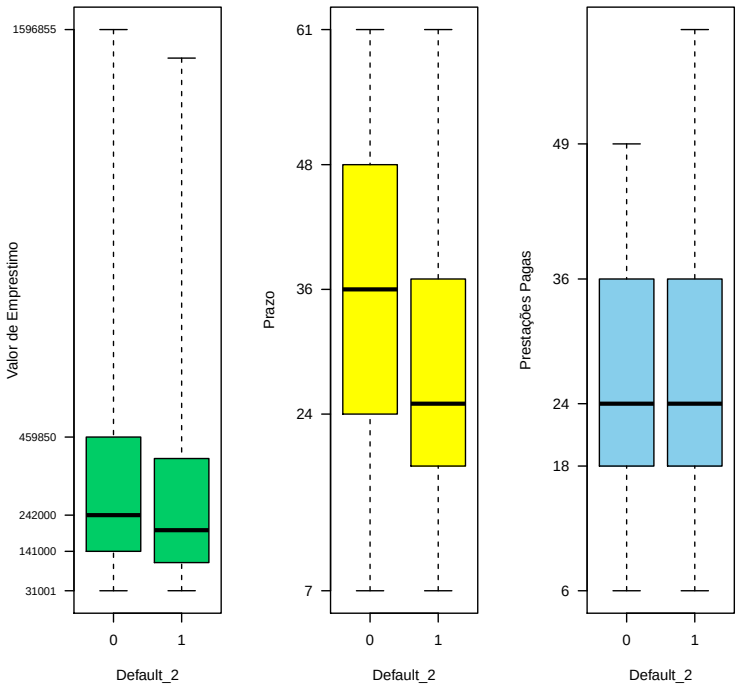


Figura 4.8: 2° Critério: Variável *default vs* variáveis Quantitativas

4.2 Probabilidade de *Default* - Regressão Logística

Nesta secção pretende-se ajustar um modelo de Regressão Logística para estimar a probabilidade de *default* de um cliente, seleccionando o melhor modelo de ajustamento. Para tal, recorrer-se-á aos métodos descritos no capítulo 3 e ao *software* $\mathcal{R}$.

4.2.1 Ajustamento dos dados - Probabilidade de *Default*

Como foram adoptados dois critérios para a definição de incumpridor, serão apresentados os respectivos modelos de ajustamento da variável resposta. De forma a não alongar a descrição dos procedimentos até se encontrar os melhores modelos em ambos os casos, todos os passos serão descritos de forma sucinta.

Em ambos os critérios é, inicialmente, considerado o seguinte conjunto de variáveis:

Idade	Agência
Género	V. Empréstimo
Estado Civil	Prazo
Habilitações	Tx. Nominal
Act. Profissional	V. Prestação
Ent. Patronal	T. Garantia

Nesta primeira análise é excluída a covariável *Prestações Pagas*, uma vez que o interesse é estudar a probabilidade de *default* para um cliente novo que pretende contratar um crédito ao consumo, pelo que, *a priori* esta informação não estará disponível.

- 1° **Critério**

Aplicando a Regressão Logística ao modelo completo, regressão descrita no capítulo anteriior, e tendo em conta a primeira definição de cliente incumpridor, ver Tabela 4.3, obteve-se os seguintes resultados da Tabela 4.6.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.5238	0.1826	-13.82	0.0000	***
Idade2	-0.0966	0.0957	-1.01	0.3128	
Idade3	-0.1974	0.0815	-2.42	0.0154	*
Idade4	-0.3453	0.0722	-4.78	0.0000	***
Idade5	-0.7440	0.0797	-9.33	0.0000	***
GeneroM	0.2611	0.0484	5.40	0.0000	***
Civil2	0.0603	0.0596	1.01	0.3118	
Habilitacoes2	-0.2270	0.0606	-3.75	0.0002	***
Habilitacoes3	-0.3593	0.0807	-4.45	0.0000	***
Habilitacoes4	-0.4999	0.1381	-3.62	0.0003	***
Habilitacoes5	-0.4186	0.1270	-3.30	0.0010	***
ActProfissional2	-0.3421	0.0703	-4.87	0.0000	***
ActProfissional3	-0.3282	0.1096	-2.99	0.0028	**
ActProfissional4	-0.6495	0.0781	-8.31	0.0000	***
EntPatronal2	-0.3671	0.0976	-3.76	0.0002	***
EntPatronal3	-0.0932	0.1118	-0.83	0.4046	
EntPatronal4	0.5847	0.1047	5.59	0.0000	***
EntPatronal5	0.9715	0.1032	9.41	0.0000	***
Agencia2	-0.0902	0.0906	-1.00	0.3196	
Agencia3	0.1209	0.1298	0.93	0.3517	
Agencia4	-0.0584	0.0783	-0.75	0.4555	
Agencia5	-0.1522	0.0817	-1.86	0.0626	.
VEmprest2	-0.2536	0.0897	-2.83	0.0047	**
VEmprest3	-0.2852	0.1068	-2.67	0.0076	**
VEmprest4	-0.2785	0.1414	-1.97	0.0488	*
Prazo2	0.5862	0.0884	6.63	0.0000	***
Prazo3	0.8989	0.0851	10.56	0.0000	***
Prazo4	1.1901	0.0915	13.01	0.0000	***
Prazo5	1.2520	0.1063	11.77	0.0000	***
TxN2	0.1068	0.0905	1.18	0.2376	
VPrest2	0.5599	0.0869	6.44	0.0000	***
VPrest3	0.5980	0.1071	5.58	0.0000	***
VPrest4	0.9666	0.1218	7.93	0.0000	***
Garantia2	-0.1707	0.0508	-3.36	0.0008	***

Tabela 4.6: Prob. *default*: Modelo Completo - 1° Critério

De seguida, realizou-se a mesma análise para o modelo nulo e, posteriormente, comparou-se os dois modelos com o propósito de escolher o modelo que melhor estima a variável resposta, através do *Teste Razão de Verosimilhanças*, descrito no capítulo anterior e obteve-se os seguintes valores:

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)	
1	22043	15118.27				
2	22010	13605.87	33	1512.41	< 2.2e-16	***

Do teste apresentado anteriormente, observa-se *p-value* muito pequeno (inferior a 2×10^{-16}), onde sepode concluir que para os níveis usuais de significância, deve-se rejeitar a hipótese de nulidade de todos os parâmetros do modelo, ou seja, deve-se rejeitar a hipótese de que o modelo nulo é melhor que o modelo complet. Pelo que, é aceitável considerar que os coeficientes acrescidos são significativamente diferentes de zero, isto é, pelos menos um deles será estatisticamente significo na modelação da variável de interesse.

Uma outra forma de avaliar os dois modelos é observar os valores do Critério de Informação de Akaike (AIC) para cada um dos modelos. Como para o modelo completo se observou um AIC de $13.673,87$, que é inferior ao do modelo nulo ($15.120,27$), conclui-se que se rejeita a hipótese, que o modelo nulo é significativamente melhor que o modelo completo.

Observando os resultados do ajustamento do modelo completo, algumas covariáveis não são significativas, o que se pode concluir através dos elevados *p-value* obtidos no teste de *Wald*. Para se encontrar um melhor modelo de ajustamento, aplicou-se o método *Stepwise - Backward*, descrito no capítulo anterior.

Segundo o método *Stepwise*, a primeira covariável a ser retirada do modelo é a que tem maior *p-value*, portanto optou-se por juntar o quarto nível da covariável *Agência* com o primeiro nível no modelo. Seguidamente, comparou-se o modelo completo com o modelo com a modificação da covariável *Agência* através do *Teste Razão de Verosimilhanças*, para o qual se obteve um *p-value* de $0,4567$. Como o *p-value* é superior ao nível de significância de 5%, por exemplo, aceita-se a hipótese de que o modelo com a transformação da covariável *Agência* é significativamente melhor que o modelo completo. Comparando os valores de AIC, retira-se a mesma conclusão, uma vez que o modelo completo possui um AIC de $13.673,87$ e o modelo em análise um AIC de $13.672,42$, que é superior ao anterior, passando a ser este o melhor modelo de ajustamento encontrado até ao momento.

Procedendo de forma análoga relativamente às restante covariáveis, revelou-se estatísticamente mais adequado proceder à remoção das covariáveis *Civil* e *Taxa Nominal* e à agrupação de vários níveis das covariáveis. Procedendo-se ao agupamento do nível 3 ao nível 1 do covariável *Entidade Patronal*, à agregação do nível 2 e 3 ao nível 1 da covariável *Agência* e à agregação do segundo nível da covariável *Idade* ao seu primeiro nível. Tendo-se obtido o modelo final, conforme ilustrado na Tabela 4.7.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.5470	0.1342	-18.98	0.0000	***
Idade3	-0.1591	0.0730	-2.18	0.0293	*
Idade4	-0.3098	0.0615	-5.04	0.0000	***
Idade5	-0.7256	0.0656	-11.06	0.0000	***
GeneroM	0.2636	0.0481	5.48	0.0000	***
Habilitacoes2	-0.2268	0.0600	-3.78	0.0002	***
Habilitacoes3	-0.3536	0.0797	-4.44	0.0000	***
Habilitacoes4	-0.4974	0.1374	-3.62	0.0003	***
Habilitacoes5	-0.4236	0.1260	-3.36	0.0008	***
ActProfissional2	-0.3496	0.0701	-4.99	0.0000	***
ActProfissional3	-0.3369	0.1096	-3.08	0.0021	**
ActProfissional4	-0.6615	0.0778	-8.50	0.0000	***
EntPatronal2	-0.2926	0.0658	-4.45	0.0000	***
EntPatronal4	0.6640	0.0753	8.82	0.0000	***
EntPatronal5	1.0441	0.0728	14.34	0.0000	***
Agencia5	-0.1080	0.0534	-2.02	0.0432	*
VEmprest2	-0.2331	0.0869	-2.68	0.0073	***
VEmprest3	-0.2645	0.1043	-2.54	0.0112	***
VEmprest4	-0.2611	0.1392	-1.88	0.0606	.
Prazo2	0.5728	0.0881	6.50	0.0000	***
Prazo3	0.8793	0.0845	10.41	0.0000	***
Prazo4	1.1700	0.0907	12.90	0.0000	***
Prazo5	1.2292	0.1054	11.66	0.0000	***
VPrest2	0.5834	0.0850	6.86	0.0000	***
VPrest3	0.6181	0.1054	5.86	0.0000	***
VPrest4	0.9837	0.1201	8.19	0.0000	***
Garantia2	-0.1603	0.0467	-3.44	0.0006	***

Tabela 4.7: Prob. *default*: Modelo de ajustamento final - 1° Critério

Faz-se notar que se testou ainda a hipótese de agregação do quarto nível da covariável *Valor de Empréstimo*, uma vez que apresenta um *p-value* de $0,0606$ mas, ao realizar-se o teste de *Teste de Razão de Verosimilhanças*, o quarto nível da covariável *Valor de Empréstimo* revelou importância para o modelo.

- 2° **Critério**

Aplicando a Regressão Logística ao modelo completo, de forma análoga ao criério acima descrito, e considerando a segunda definição de cliente incumpridor, ver Tabela 4.3, obteve-se os seguintes resultados da Tabela 4.8.

De seguida, realizou-se a mesma análise para o modelo nulo e, posteriormente, comparou-se os dois modelos com intenção de escolher o modelo que melhor estima a variável resposta, através do *Teste Razão de Verosimilhanças*.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.2217	0.1393	-8.77	0.0000	***
Idade2	-0.1362	0.0771	-1.77	0.0773	.
Idade3	-0.1131	0.0647	-1.75	0.0807	.
Idade4	-0.2864	0.0574	-4.99	0.0000	***
Idade5	-0.3847	0.0611	-6.30	0.0000	***
GeneroM	0.2066	0.0367	5.63	0.0000	***
Civil2	0.1131	0.0453	2.50	0.0126	*
Habilitacoes2	-0.0356	0.0493	-0.72	0.4698	
Habilitacoes3	-0.0708	0.0628	-1.13	0.2594	
Habilitacoes4	0.0417	0.0947	0.44	0.6596	
Habilitacoes5	0.0567	0.0960	0.59	0.5545	
ActProfissional2	-0.1661	0.0615	-2.70	0.0069	**
ActProfissional3	-0.2452	0.0911	-2.69	0.0071	**
ActProfissional4	-0.3656	0.0658	-5.55	0.0000	***
EntPatronal2	0.1353	0.0775	1.75	0.0808	.
EntPatronal3	0.2218	0.0883	2.51	0.0120	*
EntPatronal4	0.6330	0.0864	7.32	0.0000	***
EntPatronal5	0.7626	0.0867	8.80	0.0000	***
Agencia2	-0.3546	0.0675	-5.26	0.0000	***
Agencia3	-0.3325	0.0921	-3.61	0.0003	***
Agencia4	-0.8575	0.0585	-14.65	0.0000	***
Agencia5	-1.1801	0.0633	-18.63	0.0000	***
VEmprest2	-0.6360	0.0597	-10.65	0.0000	***
VEmprest3	-1.5542	0.0805	-19.30	0.0000	***
VEmprest4	-2.5193	0.1145	-22.00	0.0000	***
Prazo2	0.7465	0.0648	11.52	0.0000	***
Prazo3	1.2134	0.0651	18.64	0.0000	***
Prazo4	1.8384	0.0726	25.34	0.0000	***
Prazo5	2.5319	0.0836	30.27	0.0000	***
TxN2	-0.3640	0.0580	-6.28	0.0000	***
VPrest2	0.7690	0.0615	12.51	0.0000	***
VPrest3	1.0105	0.0795	12.71	0.0000	***
VPrest4	1.6139	0.0946	17.06	0.0000	***
Garantia2	0.0780	0.0402	1.94	0.0523	***

Tabela 4.8: Prob.*Default*: Modelo Completo - 2° Critério

Ao se realizar o *Teste Razão de Verosimilhanças*, para testar se o modelo com menos covariáveis é significativamente melhor que o modelo com mais covariáveis, à semelhança do que feito para o critério anterior, obteve-se os seguintes resultados:

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)	
1	22043	22845.46				
2	22010	20554.07	33	2291.39	< 2.2e-16	***

Observou-se um *p-value* muito pequeno (inferior a 2×10^{-16}), logo ao nível de significância 5%, ou qualquer outro usual, pode-se afirmar que se deve rejeitar a hipótese de que o modelo nulo é melhor que o modelo completo, ou seja, é aceitável considerar que os coeficientes acrescidos são significativamente diferentes de zero. Pelo que, pelos menos um deles será estatisticamente significativo na modelação da variáel resposta.

Como já referido anteriormente, uma outra forma de avaliar os dois modelos é observar os valores do Critério de Informação de Akaike (AIC) para cada um dos modelos. Como para o modelo completo se observou um AIC de $20.622,07$, que é inferior ao do modelo nulo ($22.847,46$), conclui-se que se rejeita a hipótese que o modelo nulo é significativamente melhor que o modelo completo, ou seja, alguma das covariáveis premite explicar a variável dependente.

De forma a encontrar um melhor modelo de ajustamento para a variável resposta, *default*, foi realizada de forma análoga a mesma análise efectuada para o 1° critério e tendo-se obtido os resultados da Tabela 4.9.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.2340	0.1360	-9.08	0.0000	***
Idade2e3	-0.1144	0.0590	-1.94	0.0524	.
Idade4	-0.2795	0.0570	-4.90	0.0000	***
Idade5	-0.3761	0.0604	-6.23	0.0000	***
GeneroM	0.2064	0.0366	5.64	0.0000	***
Civil2	0.1118	0.0452	2.47	0.0134	*
ActProfissional2	-0.1808	0.0583	-3.10	0.0019	**
ActProfissional3	-0.2408	0.0864	-2.79	0.0053	**
ActProfissional4	-0.3841	0.0602	-6.38	0.0000	***
EntPatronal2	0.1315	0.0773	1.70	0.0889	.
EntPatronal3	0.2150	0.0880	2.44	0.0145	*
EntPatronal4	0.6266	0.0860	7.28	0.0000	***
EntPatronal5	0.7574	0.0864	8.76	0.0000	***
Agencia2	-0.3583	0.0673	-5.32	0.0000	***
Agencia3	-0.3373	0.0918	-3.67	0.0002	***
Agencia4	-0.8533	0.0583	-14.63	0.0000	***
Agencia5	-1.1828	0.0633	-18.69	0.0000	***
VEmprest2	-0.6332	0.0597	-10.61	0.0000	***
VEmprest3	-1.5524	0.0805	-19.29	0.0000	***
VEmprest4	-2.5179	0.1144	-22.00	0.0000	***
Prazo2	0.7451	0.0647	11.51	0.0000	***
Prazo3	1.2114	0.0650	18.64	0.0000	***
Prazo4	1.8366	0.0723	25.41	0.0000	***
Prazo5	2.5288	0.0833	30.35	0.0000	***
Tx2	-0.3668	0.0576	-6.36	0.0000	***
VPrest2	0.7691	0.0614	12.53	0.0000	***
VPrest3	1.0119	0.0793	12.76	0.0000	***
VPrest4	1.6199	0.0942	17.19	0.0000	***
Garantia2	0.0762	0.0402	1.90	0.0577	.

Tabela 4.9: Prob. *default*: Modelo de ajustamento final - 2° Critério

Testou-se ainda a hipótese de juntar as covariáveis com o *p-value* superior ao nível base, mas tais transformações demonstraram que o modelo se tornaria menos explicativo para a variável resposta, pelo que se optou incorporar as covariáveis no modelo.

4.2.2 Análise dos resíduos

Concluído o ajustamento dos dados em relação à probabilidade de um cliente ser incumpridor, deve-se testar a adequabilidade dos modelos encontrados e esta é verificada através da análise dos resíduos.

- 1° **Critério**

A análise de resíduos foi efectuada com base nos desvios residuais, concluindo-se que os pontos apresentam um desvio residual padronizado de valor compreendido entre $[-2, 3]$, como se pode observar na Figura 4.9. E ainda, se observa que os valores dos resíduos são iguais em magnitude e opostos em sinal para cada par de valores na mesma categoria, consequência da equação (3.33).

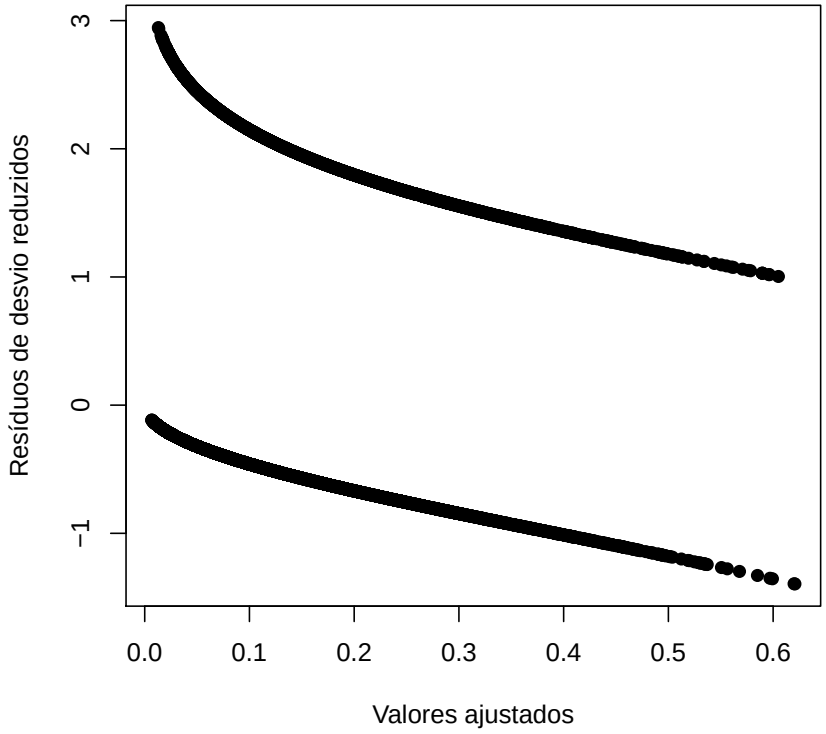


Figura 4.9: Prob. *default*: Desvios residuais reduzidos - 1° Critério

Na análise das observações com repercussão elevada, optou-se por considerar que são observações mesmo discordantes as observações tais que $\frac{nh_{ii}}{p} > 5,5$, pelo que se optou por excluir tais observações do modelo de ajustamento.

Quanto à análise de pontos influentes, recorreu-se à distância de Cook e considerou-se observações influentes aquelas cuja distância de Cook é superior a $0,001$, por serem as observações com valores de distância de Cook mais elevados, tendo-se removido tais observações do modelo de ajustamento.

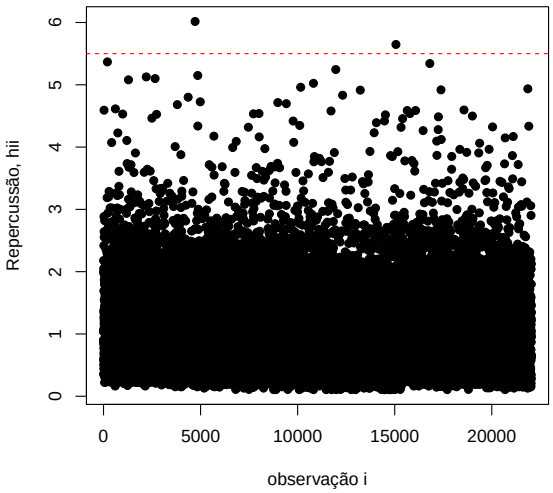


Figura 4.10: Prob. *default*: Observações com repercussão Elevada - 1° Critério

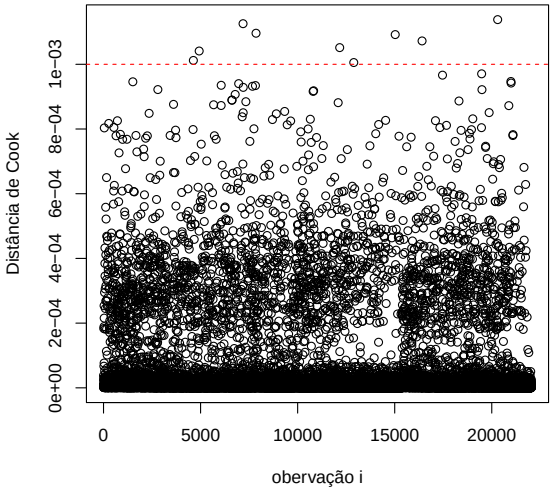


Figura 4.11: Prob. *default*: Distâncias de Cook - 1° Critério

- 2° **Critério**

De forma análoga, a análise de resíduos foi feita com base nos desvios residuais, concluindo-se que os pontos apresentam um desvio residual padronizado de valor compreendido entre $[-2, 3]$, como se pode observar na Figura 4.12 e conclusão é semelhante à do primeiro critério, observa-se que para cada categoria da variável resposta os valores dos resíduos serão de magnitude e opostos em sinal para cada par de valores na mesma categoria.

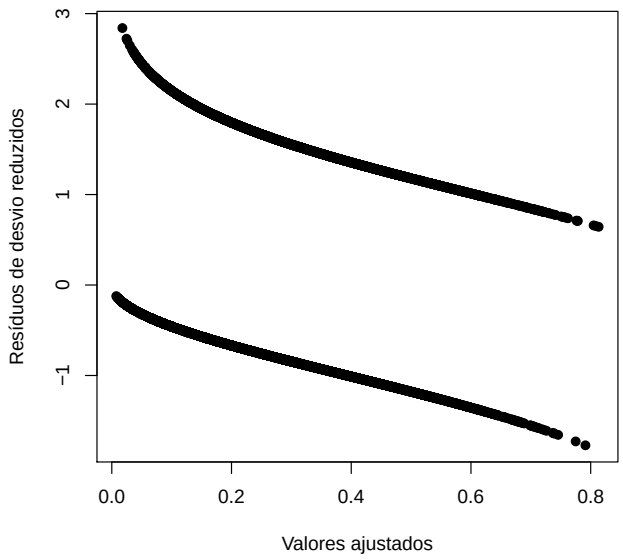


Figura 4.12: Prob. *default*: Desvios residuais reduzidos - 2° Critério

Na análise das observações com repercussão elevada, optou-se por considerar que são observações mesmo discordantes as observações tais que $\frac{nh_{ii}}{p} > 3,4$, pelo que se excluiu as observações consideradas com repercurssão elevada do modelo de ajustamento final.

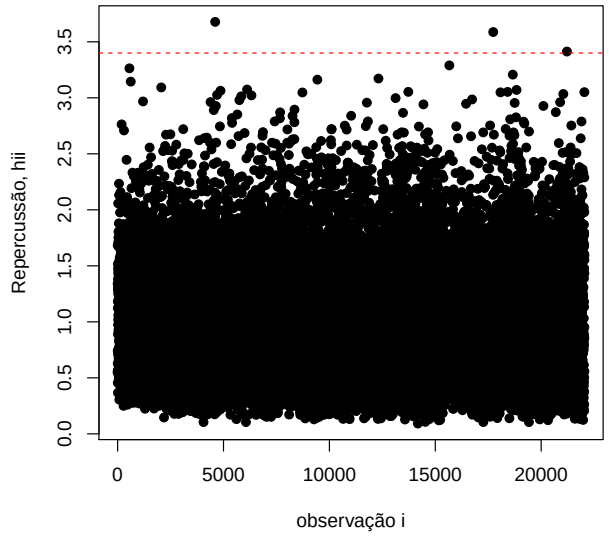


Figura 4.13: Prob. *default*: Observações com repercussão Elevada - 2° Critério

Quanto à análise de pontos influentes, recorreu-se à distância de Cook e considerou-se obervações influentes tais que a distância de Cook é superior a $0,00045$, por serem as observações com valores de distâcia de Cook mais elevados. Logo, foram removidas as observações consideradas como influentes do modelo de ajustamento.

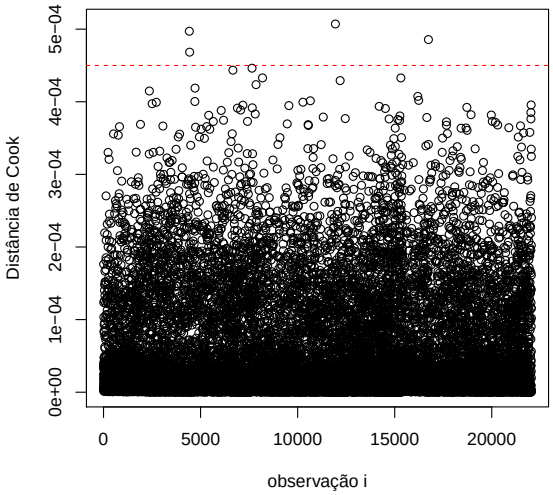


Figura 4.14: Prob. *default*: Distâncias de Cook - 2° Critério

4.2.3 Estimação da probabilidade de *Default*

Após o estudo estatístico do melhor modelo para ambos os critérios, a análise de resíduos respectiva e a exclusão das observações discordantes, pode-se apresentar os resultados finais da estimação da probabilidade de *default* de cada cliente, π_i, definida no capítulo anterior e que pode ser escrita da seguinte forma:

$$\pi_i = \frac{e^{Y_i}}{1+e^{Y_i}}$$

onde, $Y_i = \beta_0 + \beta_1 X_1 + ... \beta_p X_p$ com $Y_i \sim B(1, \pi_i)$, $(\beta_0, \beta_1, ..., \beta_p)^T$ parâmetros de regressão e **X** representa o vector de covariáveis.

Para cada um dos critérios define-se como cliente padrão aquele cujas características se encontrem agrupadas nos níveis da base de todas as covariáveis incluídas no modelo. Assim, o cliente padrão possui as seguintes características, relativamente ao dois critérios:

1° **Critério**		2° **Critério**	
Variável	Grupo	Variável	Grupo
Idade1e2	Inferior a 29	Idade1	Inferior a 26
GeneroF	Feminino	GeneroF	Feminino
Habilitacoes1	Habilitações Desconhecidas	Civil1	Casado, Divorciado, Separado
ActProfissional1	Peq./Méd. Empresa e Outros	ActProfissional1	Peq./Méd. Empresa e Outros
EntPatronal1	Inst. Financeiras e Outros	EntPatronal1	Inst. Financeiras e Outros
Agencia1,2,3e4	<> (3, 12, 19, 22, 25, 27, 31, 32)	Agencia1	4, 7, 14, 18, 28 e 29
VEmprest1	≤ 107.600	VEmprest1	≤ 107.600
Prazo1	Inferior a 13	Prazo1	Inferior a 13
VPrest1	≤ 6.120	Tx1	< 12,5
Garantia1	Depósitos, Hipotecas e Outros	VPrest1	≤ 6.120
		Garantia1	Depósitos, Hipotecas e Outros

Tabela 4.10: Prob. *default*: Cliente Padrão

Tendo como cliente padrão, o cliente que assume as características acima descritas, estimou-se que a probabilidade de um cliente padrão entrar em incumprimento, para cada um dos critérios é de:

	π_i
1° Critério	0,072
2° Critério	0,222

Como se observa, a probabilidade de *default* do 2° Critério é superior, mas como as covariáveis do cliente padrão não são as mesmas nos dois critérios, não é correcto fazer comparações, uma vez que poderá ter influências nos resultados observados.

Na Tabela 4.11 apresenta-se para cada, um dos critérios adoptados os valores estimados para a probabilidade de um cliente ser incumpridor para cada uma das características.

1° **Critério**		2° **Critério**	
Características	Coeficientes	Características	Coeficientes
Cliente Padrão	-2.5591	Cliente Padrão	-1.2517
Idade3	-0.1629	Idade2e3	-0.1128
Idade4	-0.3136	Idade4	-0.2793
IdadeIDA5	-0.7334	Idade5	-0.3772
GeneroM	0.2619	GeneroM	0.2078
Habilitacoes2	-0.2268	Civil2	0.1109
Habilitacoes3	-0.3536	ActProfissional2	-0.1793
Habilitacoes4	-0.6471	ActProfissional3	-0.2433
Habilitacoes5	-0.4077	ActProfissional4	-0.3836
ActProfissional2	-0.3509	EntPatronal2	0.1388
ActProfissional3	-0.3779	EntPatronal3	0.2232
ActProfissional4	-0.6631	EntPatronal4	0.6322
EntPatronal2	-0.2794	EntPatronal5	0.7641
EntPatronal4	0.6735	Agencia2	-0.3564
EntPatronal5	10.560	Agencia3	-0.3300
Agencia5	-0.1107	Agencia4	-0.8532
VEmprest2	-0.2517	Agencia5	-1.1838
VEmprest3	-0.2704	VEmprest2	-0.6355
VEmprest4	-0.2629	VEmprest3	-1.5621
Prazo2	0.5712	VEmprest4	-2.5414
Prazo3	0.8870	Prazo2	0.7549
Prazo4	11.774	Prazo3	1.2229
Prazo5	12.388	Prazo4	1.8510
VPrest2	0.5955	Prazo5	2.5459
VPrest3	0.6324	TxN2	-0.3682
VPrest4	0.9949	VPrest2	0.7755
Garantia2	-0.1570	VPrest3	1.0196
		VPrest4	1.6338
		Garantia2	0.0766

Tabela 4.11: Prob. *default*: Ajustamento da Probabilidade de *default*

Ilustram-se seguidamente, na Tabela 4.12, alguns exemplos de clientes com características diferentes, de acordo com os variáveis que constituem os modelos de estimação.

	Níveis das covariáveis											
Cliente	Idade	Genero	Civil	Habilitacoes	ActProf.	EntPatronal	Agencia	VEmprest	Prazo	VPrest	TxN	Garantia
a	1	F	2	1	2	4	4	3	4	2	3	2
b	4	M	2	1	1	4	5	3	4	2	3	2
c	4	M	2	2	1	4	3	1	1	2	1	2
d	4	F	2	1	1	2	4	1	4	1	1	2
e	5	M	2	2	2	4	1	4	5	2	4	1

Tabela 4.12: Clientes Ilustrativos

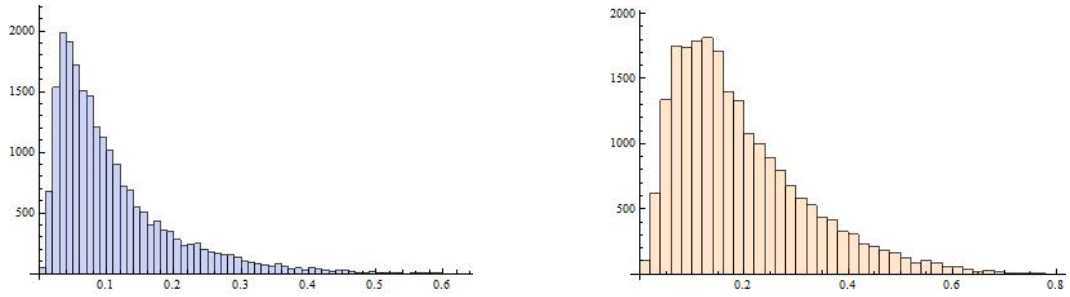
Tendo-de obtido as seguintes estimativas para os clientes acima desctritos, tendo em conta os critérios inicialmente escolhidos, de:

	1° Critério		2° Critério	
Cliente	y_i	π_i	y_i	π_i
a	-0,8693	0,2954	-1,7313	0,1504
b	-0,6739	0,3376	-1,9541	0,1241
c	-2,3163	0,0898	-1,3892	0,1995
d	-2,1397	0,1053	-0,3944	0,4027
e	-0,9696	0,2750	-1,3319	0,2088

Tabela 4.13: Prob. *default*: Exemplos

Estimou-se, ainda, a probabilidade de *default* da carteira de crédito ao consumo da instituição bancária de Cabo Verde, segundo os dois critérios definidos. Na Figura 4.15 pode-se observar a distribuição da probabilidade de *default* da carteira, para ambos os critérios.

	π_i
1° Critério	0,108287
2° Critério	0,193863



(a) 1° Critério (b) 2° Critério

Figura 4.15: Distribuição da probabilidade de *default* da carteria

Uma outra análise que se pode realizar aos resultados está relacionada com a *Odds Ratio*, pelo que se apresentam algumas análises de exemplo.

	Odds Ratio	
	1° Critério	2° Critério
Feminino *vs* Masculino	0,8511016	0,854099644
Solteiro *vs* Casado	1	1,088880295
35 anos *vs* 45 anos de idade	1,487570659	1,084560443
Prazo 24 meses *vs* 48 meses	0,605180043	0,58604994

Tabela 4.14: Prob. *default*: *Odds Ratio*

Os valores acima indicam por exemplo, para o primeiro critério, que a probabilidade de um cliente do género Feminino vir a ser incumpridor é 0,851 vezes a probabilidade de um cliente do género Masculino, mantendo rudo o resto constante, ou seja, a probabilidade de incumprimento é menor entre clientes do género Feminino do que entre clientes do género Masculino. Também se pode observar que a probabilidade de um cliente de crédito com 35 anos de idade vir a ser incumpridor é aproximadamente superior 1,488 vezes à de um cliente com 45 anos de idade.

4.3 Proporção das Prestações Pagas - Regressão Beta

Uma vez estimada a probabilidade de *default* para um novo contrato de crédito ao consumo, considerar-se-á relevante efectuar uma análise complementar que permite estimar a *Proporção de Prestações Pagas* de um cliente classificado como incumpridor, considerando os dois critérios adoptados para a definição de incumpridor (ver Tabela 4.3). Como facilmente se compreende a de proporção de prestações pagas está relacionada com a taxa de recuperação do empréstimo concedido, pelo que se torna relevante efectuar a sua avaliação.

Considerou-se importante, de um ponto vista prático, estudar a proporção de Prestações Pagas no final do contrato, porque segundo o 1° critério de incumpridor, apesar de, algures durante o contrato, um cliente ter estado pelo menos 90 dias com prestações em atraso, alguns destes chegam ao termo do contrato com todas as prestações pagas.

Por outro lado, há clientes que são considerados incumpridores por terem dívidas no final do contrato (segundo critério), mas que não cumprem o primeiro critério de incumprimento uma vez que não estiveram durante o contrato mais do que 90 dias sem efectuar pagamentos de prestações.

É ainda importante ainda analisar se o número de prestações em dívida, à data de encerramento do contrato é elevado ou residual, como indicador do montante de capital recuperado até ao final do contrato, em caso de ocorrência de *default*.

Assim nesta secção utilizar-se-á a Regressão Beta para estimar a proporção de prestações pagas, uma vez que esta variável toma valores entre 0 e 1 e utilizando métodos descritos no capítulo 3 e recorrendo ao *software* $\mathcal{R}$.

4.3.1 Ajustamento dos dados - Proporção de Prestações Pagas

Note-se que a variável em estudo, a proporção de prestações pagas, assume por vezes valor 0, quando nenhuma prestação foi liquidada e toma também o valor 1, se o cliente chegou ao final do processo com todas as prestações pagas. Desta forma, seguindo a sugestão de [FCN04], efectuou-se a mudança da variável,

$$w_i = \frac{y_i(n-1)+0.5}{n}.$$

Como foram adoptados dois critérios para a definição de cliente incumpridor, serão apresentados os respectivos modelos de ajustamento da variável resposta.

Em ambos os critérios, é inicialmente considerado o seguinte conjunto de variáveis:

Idade	Agência
Género	V. Empréstimo
Estado Civil	Prazo
Habilitações	Tx. Nominal
Act. Profissional	V. Prestação
Ent. Patronal	T. Garantia

- 1° **Critério**

Antes de se iniciar a análise, note-se que foram excluídas 27 observações, uma vez que tomam valores superiores a 1. Estas situações, de acordo com a instituição bancária, reportam-se a clientes que, durante o contrato, dispuseram de mais prestações para finalizar o pagamento da dívida. Assim, para efeitos deste estudo, foram utilizadas $98,87\%$ das observações da base de dados.

Ajustanto uma Regressão Beta ao modelo completo e sendo a proporção das prestações pagas a variável resposta, com a mudança de variável atrás definida, e tendo em conta a primeira definição de cliente incumpridor, ver Tabela 4.3, obteve-se os seguintes resultados:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.810822	0.180142	15.603	< 2e-16	***
TxN2	-0.453035	0.088225	-5.135	2.82e-07	***
VPrest2	0.510818	0.082757	6.172	6.72e-10	***
VPrest3	0.748608	0.101881	7.348	2.01e-13	***
VPrest4	1.210366	0.117818	10.273	< 2e-16	***
VEmprest2	-0.369884	0.088383	-4.185	2.85e-05	***
VEmprest3	-0.967406	0.104758	-9.235	< 2e-16	***
VEmprest4	-1.743209	0.141660	-12.306	< 2e-16	***
Idade2	0.009950	0.089738	0.111	0.91171	
Idade3	-0.024629	0.076594	-0.322	0.74779	
Idade4	0.018514	0.068726	0.269	0.78763	
Idade5	0.100167	0.075777	1.322	0.18621	
Prazo2	0.199566	0.088982	2.243	0.02491	*
Prazo3	0.434830	0.088922	4.890	1.01e-06	***
Prazo4	0.635942	0.094850	6.705	2.02e-11	***
Prazo5	1.120210	0.109306	10.248	< 2e-16	***
Agencia2	-0.044063	0.084365	-0.522	0.60147	
Agencia3	0.290362	0.126614	2.293	0.02183	*
Agencia4	-0.519201	0.075392	-6.887	5.71e-12	***
Agencia5	-0.857884	0.077859	-11.018	< 2e-16	***
ActProfissional2	-0.005623	0.066192	-0.085	0.93230	
ActProfissional3	-0.089133	0.106382	-0.838	0.40211	
ActProfissional4	-0.124209	0.075243	-1.651	0.09878	.
GeneroM	0.023298	0.046832	0.497	0.61886	
EntPatronal2	0.070722	0.096517	0.733	0.46372	
EntPatronal3	0.205146	0.109971	1.865	0.06212	.
EntPatronal4	0.295649	0.102419	2.887	0.00389	**
EntPatronal5	0.262633	0.099851	2.630	0.00853	**
Civil2	-0.103437	0.057516	-1.798	0.07211	.
Habilitacoes2	-0.075729	0.058520	-1.294	0.19564	
Habilitacoes3	-0.112961	0.079452	-1.422	0.15510	
Habilitacoes4	-0.184407	0.136398	-1.352	0.17638	
Habilitacoes5	-0.119181	0.125935	-0.946	0.34396	
Garantia2	0.077494	0.049609	1.562	0.11826	

Tabela 4.15: PPagas: Modelo Completo - 1° Critério

Em seguida, realizou-se a mesma análise para o modelo nulo e posteriormente, comparou-se os dois modelos com o propósito de escolher o modelo que melhor estima a variável resposta. Para se comparar os modelos, com o objectivo de seleccionar o melhor modelo, observou-se os valores do Critério de Informação de Akaike (AIC), sendo o valor de AIC para o modelo completo de $-14.362,05$ e para o modelo nulo de $-1930,61$. Como este último é superior, logo o modelo que melhor define a variável resposta é o modelo completo, ou seja, pode-se afirmar que, é aceitável considerar que os coeficientes acrescidos são significativamente diferentes de zero, donde se concluir que pelo menos um dos coeficientes é estatisticamente signficativo na modelação da variável de interesse.

Observando os resultados do ajustamento do modelo completo, algumas covariáveis não são significativas, o que se pode concluir através dos elevados valores de *p-value* observados.

À semelhança do que feito para a Regressão Logística, aplicou-se o método *Stepwise - Backward*, para seleccionar as covariáveis.

Segundo o método *Stepwise*, a primeira covariável a ser retirada do modelo é a que tem maior *p-value*, logo optou-se por agrupar o segundo e, posterior-mente, o terceiro nível da covariável *Actividade Profissional* ao nível do cliente padrão, obtendo-se um AIC de $-14.365,251$, que é inferior ao valor de AIC do modelo completo. Considere-se, assim, que o modelo com a transformação da covariável *Actividade Profissional* se revelou um ajustamento estatisticamente mais adequado.

De forma a não alongar a descrição dos procedimentos até se encontrar um bom modelo, os passos seguintes foram semelhantes aos descritos acima, identificando a covariável com um *p-value* mais elevado e, posteriormente, comparando os modelos através do valor AIC. O modelo final de ajustamento encontrado para a variável reposta, proporção das prestações pagas, segundo o 1° critério de cliente incumpridor, é apresentado na Tabela 4.16

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.86599	0.14435	19.854	< 2e-16	***
TxN2	-0.44857	0.08771	-5.114	3.15e-07	***
VPrest2	0.51734	0.08262	6.262	3.80e-10	***
VPrest3	0.75342	0.10152	7.422	1.16e-13	***
VPrest4	1.21266	0.11737	10.332	< 2e-16	***
VEmprest2	-0.37937	0.08814	-4.304	1.67e-05	***
VEmprest3	-0.97605	0.10428	-9.359	< 2e-16	***
VEmprest4	-1.74345	0.14120	-12.347	< 2e-16	***
Idade5	0.09910	0.05359	1.849	0.064393	.
Prazo2	0.19934	0.08881	2.245	0.024795	*
Prazo3	0.43838	0.08853	4.952	7.36e-07	***
Prazo4	0.63129	0.09423	6.699	2.10e-11	***
Prazo5	1.11395	0.10839	10.278	< 2e-16	***
Agencia3	0.37412	0.11270	3.320	0.000901	***
Agencia4	-0.45156	0.05483	-8.236	< 2e-16	***
Agencia5	-0.80937	0.06109	-13.250	< 2e-16	***
ActProfissional4	-0.10586	0.05201	-2.035	0.041829	*
EntPatronal3	0.14378	0.07426	1.936	0.052858	.
EntPatronal4	0.24506	0.06164	3.976	7.02e-05	***
EntPatronal5	0.20775	0.05749	3.614	0.000301	***
Civil2	-0.10270	0.05647	-1.819	0.068946	.
Habilitacoes2e3	-0.08948	0.05184	-1.726	0.084368	.
Habilitacoes4e5	-0.17963	0.09365	-1.918	0.055110	.

Tabela 4.16: PPagas: Modelo final de ajustamento - 1° Critério

Testou-se, ainda, a agregação das covariáveis com o *p-value* elevado ao nível base, ou a outros níveis, ou até mesmo exclui-las, mas tais transformções produziram um modelo menos explicativo da variável resposta, pelo que se optou incorpar as covariáveis no modelo.

- 2° **Critério**

Antes de se iniciar a análise, note-se que foram excluídas 33 observações, uma vez que tomam valores superiores a 1, como referido anteriormente. Assim para efeitos de estudo foram usadas $99,28\%$ observações da base de dados de clientes.

Ajustando uma Regressão Beta ao modelo completo, e sendo a proporção das prestações pagas a variável resposta, tendo em conta o segundo critério definido para cliente incumpridor, ver Tabela 4.3, obteve-se os seguintes resultados:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.628800	0.131305	20.021	<2e-16	***
TxN2	-0.146266	0.052111	-2.807	0.00500	**
VPrest2	0.085661	0.057273	1.496	0.13474	
VPrest3	0.105079	0.073842	1.423	0.15473	
VPrest4	0.156531	0.090899	1.722	0.08507	.
VEmprest2	0.042232	0.053481	0.790	0.42973	
VEmprest3	-0.044457	0.075189	-0.591	0.55434	
VEmprest4	0.019712	0.110691	0.178	0.85866	
Idade2	0.004386	0.066947	0.066	0.94777	
Idade3	-0.083129	0.056255	-1.478	0.13948	
Idade4	-0.079056	0.050274	-1.573	0.11583	
Idade5	-0.125498	0.053897	-2.328	0.01989	*
Prazo2	-0.077342	0.061006	-1.268	0.20488	
Prazo3	-0.053020	0.062629	-0.847	0.39723	
Prazo4	-0.313327	0.069885	-4.483	7.34e-06	***
Prazo5	-0.437777	0.079771	-5.488	4.07e-08	***
Agencia2	0.385438	0.056128	6.867	6.55e-12	***
Agencia3	0.629003	0.079189	7.943	1.97e-15	***
Agencia4	0.382008	0.048093	7.943	1.97e-15	***
Agencia5	0.106331	0.055121	1.929	0.05372	.
ActProfissional2	0.082389	0.053058	1.553	0.12047	
ActProfissional3	0.200515	0.081656	2.456	0.01407	*
ActProfissional4	0.020616	0.058045	0.355	0.72246	
GeneroM	0.010785	0.033030	0.327	0.74403	
EntPatronal2	0.032279	0.073957	0.436	0.66250	
EntPatronal3	0.030443	0.083274	0.366	0.71468	
EntPatronal4	-0.041274	0.080383	-0.513	0.60762	
EntPatronal5	0.039873	0.080332	0.496	0.61964	
Civil2	-0.088837	0.041976	-2.116	0.03431	*
Habilitacoes2	-0.135154	0.043709	-3.092	0.00199	**
Habilitacoes3	-0.177442	0.056180	-3.158	0.00159	**
Habilitacoes4	-0.235959	0.085794	-2.750	0.00595	**
Habilitacoes5	-0.404036	0.088484	-4.566	4.97e-06	***
Garantia2	-0.308352	0.035473	-8.693	< 2e-16	***

Tabela 4.17: PPagas: Modelo Completo - 2° Critério

De seguida, analogamente aos ajustamentos descritos analogamente, realizou-se a mesma análise para o modelo nulo e posteriormente, comparou-se os dois modelos com o propósito de escolher o modelo que melhor estima a variável resposta. Pelo que, se optou pelo modelo completo, uma vez que este se revelou ser estatisticamente mais significativo que o modelo nulo.

Seguindo a analogia apresentada nos casos anteriores, utilizando o método *Stepwise* para seleccionar as covariáveis a retirar do modelo e o valor AIC para comprar os modelos, tendo em conta o 2° critério de cliente incumpridor, obteve-se o melhor modelo de ajustamento para a variável em estudo, a proporção das prestações pagas (Tabela 4.18).

Note-se que, ainda, se agrupou os níveis das covariáveis com um *p-value* mais elevado ao nível base, mas ao realizar o *Teste de Razão de Verosimilhanças*, mas tais transformções demonstraram um modelo menos explicativo da variável resposta, pelo que se optou incorpar as covariáveis no modelo.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.61160	0.08795	29.693	< 2e-16	***
TxN2	-0.14251	0.05035	-2.831	0.00465	**
VPrest3	0.09964	0.04708	2.117	0.03430	*
VPrest4	0.15611	0.05129	3.044	0.00234	**
VEmprest3	-0.07436	0.03606	-2.062	0.03922	*
Idade4	-0.07842	0.04018	-1.952	0.05097	.
Idade5	-0.12688	0.04768	-2.661	0.00779	**
Prazo4	-0.24877	0.03838	-6.482	9.05e-11	***
Prazo5	-0.37183	0.04546	-8.179	2.85e-16	***
Agencia2	0.39358	0.05472	7.193	6.33e-13	***
Agencia3	0.64105	0.07804	8.214	< 2e-16	***
Agencia4	0.39370	0.04730	8.324	< 2e-16	***
Agencia5	0.10942	0.05478	1.998	0.04575	*
ActProfissional2	0.05798	0.03397	1.707	0.08783	.
ActProfissional3	0.17679	0.06825	2.590	0.00959	**
Civil2	-0.08541	0.04162	-2.052	0.04016	*
Habilitacoes2	-0.12258	0.03987	-3.075	0.00211	**
Habilitacoes3	-0.15524	0.05234	-2.966	0.00302	**
Habilitacoes4	-0.20641	0.08271	-2.496	0.01258	*
Habilitacoes5	-0.37848	0.08587	-4.408	1.05e-05	***
Garantia2	-0.31282	0.03522	-8.881	< 2e-16	***

Tabela 4.18: PPagas: Modelo final de ajustamento - 2° Critério

4.3.2 Análise de Resíduos

Finalizado o ajustamento dos dados em relação à *Proporção das Prestações Pagas*, deve-se testar a adequabilidade dos modelos encontrados, pelo que é necessário realizar a análise dos resíduos.

- 1° **Critério**

A análise de resíduos foi efectuada com base nos desvios residuais, concluindo-se que os pontos apresentam um desvio residual padronizado de valor compreendido entre $[-2; 5, 5]$, como se pode observar na Figura 4.16, apenas se destacam duas observações com valores residuais mais elevados.

Quanto à análise de pontos influentes, recorreu-se à distância de Cook e considerou-se obervações influentes aquelas cuja distância de Cook é superior a $0, 010$, sendo as distâncias que se destacam. De forma a encontrar um melhor modelo de ajustamento da proporção das prestações pagas, procedeu-se à remoção das observações consideradas como influentes.

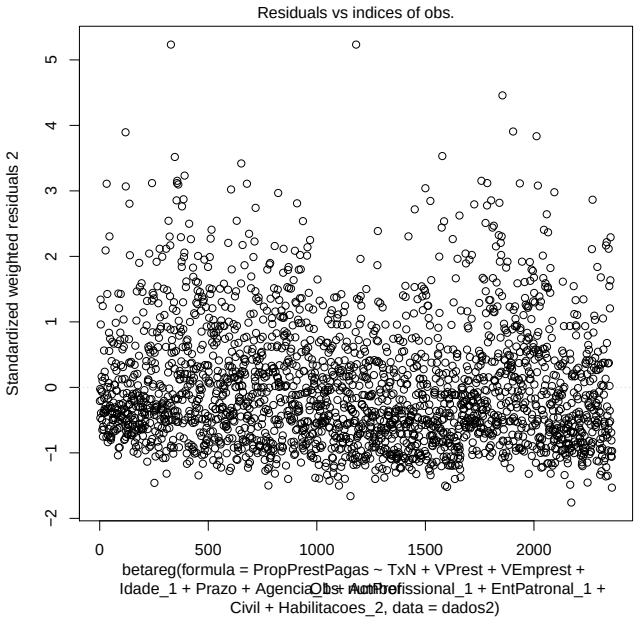


Figura 4.16: PPagas: Análise de Resíduos - 1° Critério

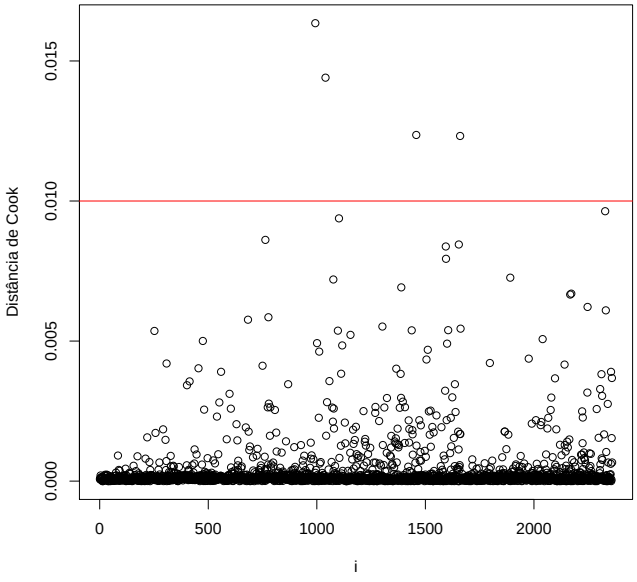


Figura 4.17: PPagas: Distâncias de Cook - 1° Critério

- 2° **Critério**

De forma análoga, a análise de resíduos foi realizada com base nos desvios residuais, concluindo-se que os pontos apresentam um desvio residual padronizado de valor compreendido entre $[-2, 2]$, como se pode observar na figura 4.18.

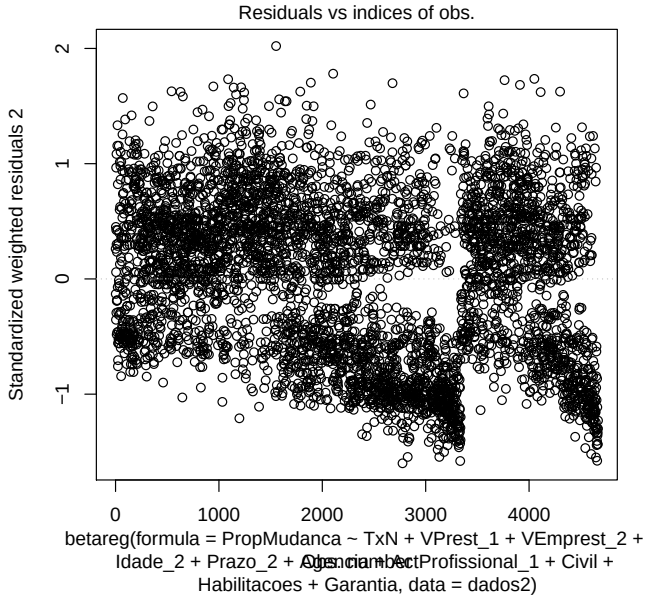


Figura 4.18: PPagas: Análise de Resíduos - 2° Critério

Quanto à análise de pontos influentes, recorreu-se à distância de Cook e considerou-se obervações influentes aquelas cuja distância de Cook é superior a $0,008$, sendo as distâncias que mais se destacam. Tendo sido exclúidas as observações consideradas como influentes.

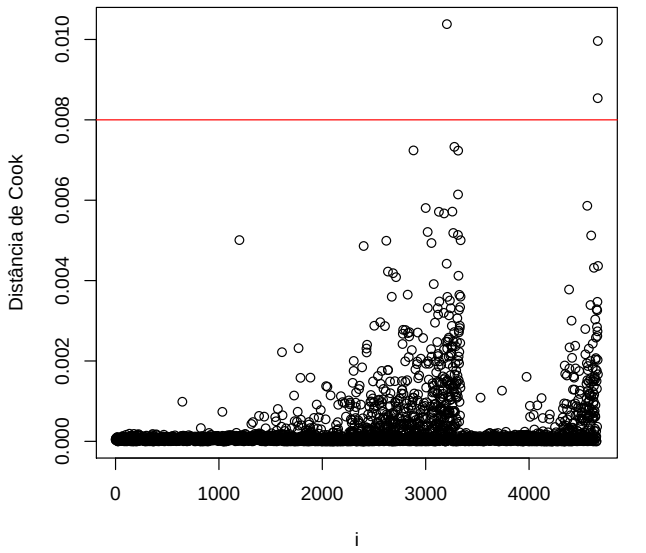


Figura 4.19: PPagas: Distâncias de Cook - 2° Critério

4.3.3 Estimação da Proporção das Prestações Pagas

Concluido o estudo estatístico do um modelo adequado para ambos os critérios, a análise de resíduos respectiva e a exclusão das observações discordantes, pode-se apresentar os resultados obtidos da estimação da *Proporção das Prestações Pagas* de um dado cliente definido como incumpridor, que se designa como b_i. E que pode ser escrita da seguinte forma:

$$b_i = \frac{exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1+exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}$$

onde, $Y_i = \beta_0 + \beta_1 X_1 + ...\beta_p X_p$ com $Y \sim \mathcal{B}(\mu, \phi)$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^T$ vector de parâmetros de regressão desconhecidos e $\boldsymbol{x_i} = (x_{i1}, ..., x_{ik})^T$ representa o vector de covariáveis.

Para cada um dos critérios define-se como cliente padrão aquele cujas características se encontrem agrupadas nos níveis da base de todas as covariáveis incluídas no modelo. Assim, o cliente padrão possui as seguintes características, relativamente ao dois critérios:

1° Critério		**2° Critério**	
Variável	Grupo	Variável	Grupo
Idade1,2,e4	Inferior a 29 e Entre 33 e 42	Idade1,2e3	Inferior a 33
Civil1	Casado, Divorciado, Separado	Civil1	Casado, Divorciado, Separado
Habilitacoes1	Habilitações Desconhecidas	Habilitacoes1	Habilitações Desconhecidas
ActProfissional1,2e4	Doméstica, Peq./Méd. Empresa e Outros	ActProfissional1,2e4	Doméstica, Peq./Méd. Empresa, Emp. Escriório e Outros
EntPatronal1	Inst.Financeiros e Outros	Agencia1	4, 7, 14, 18, 28 e 29
Agencia1	4, 7, 14, 18, 28 e 29	VEmprest1,2e4	<200.000 e $\geq$ 512.320
VEmprest1	$\leq$ 107.600	Prazo1,2e3	Inferior a 37
Prazo1	Inferior a 13	VPrest1,2e3	$\leq$ 6.120
VPrest1	$\leq$ 6.120	Tx1	<12,5
TxN1	<12,5	Garantia1	Depósitos e Hipotecas

Tabela 4.19: PPagas: Cliente Padrão

Tendo como cliente padrão, o cliente que assume as características acima descritas, desfazendo a mudança de variável, estimou-se que a proporção das prestações pagas, para cada um dos critérios é de:

	b_i
1° Critério	0,9462107
2° Critério	0,9316240

Como se observa, a proporção das prestações pagas dos dois os critérios adoptados para definir cliente incumpridor, são muito semelhantes e de valor bastante elevado. Como se pode verificar estimou-se que para um cliente padrão, com as características acima definidas, classificado como incumpridor ao longo do contrato, segundo o $1°$ critério, $94,62\%$ das prestações encontram-se pagas, a mesma estimativa será de $93,3\%$ se se optar pelo $2°$ critério de classificação de cliente incumpridor.

Seguidamente apresenta-se, para cada um dos critérios adoptados, os valores estimados para a *Proporção das Prestações Pagas* de um cliente incumpridor para cada uma das características , desfazendo a mudança de variável.

1° **Critério**		2° **Critério**	
Características	Coeficientes	Características	Coeficientes
Cliente Padrão	2.86699	Cliente Padrão	2.61160
TxN2	-0.44897	TxN2	-0.14251
VPrest2	0.51735	VPrest3	0.09964
VPrest3	0.75353	VPrest4	0.15611
VPrest4	1.21296	VEmprest3	-0.07436
VEmprest2	-0.37974	Idade4	-0.07842
VEmprest3	-0.97668	Idade5	-0.12688
VEmprest4	-1.74440	Prazo4	-0.24877
Idade5	0.09893	Prazo5	-0.37183
Prazo2	0.19921	Agencia2	0.39358
Prazo3	0.43835	Agencia3	0.64105
Prazo4	0.631355	Agencia4	0.39370
Prazo5	1.11421	Agencia5	0.10942
Agencia3	0.37406	ActProfissional2	0.05798
Agencia4	-0.45196	ActProfissional3	0.17679
Agencia5	-0.80992	Civil2	-0.08541
ActProfissional4	-0.10612	Habilitacoes2	-0.12258
EntPatronal3	0.14363	Habilitacoes3	-0.15524
EntPatronal4	0.24495	Habilitacoes4	-0.20641
EntPatronal5	0.20763	Habilitacoes5	-0.37848
Civil2	-0.10296	Garantia2	-0.31282
Habilitacoes2e3	-0.08973		
Habilitacoes4e5	-0.17992		

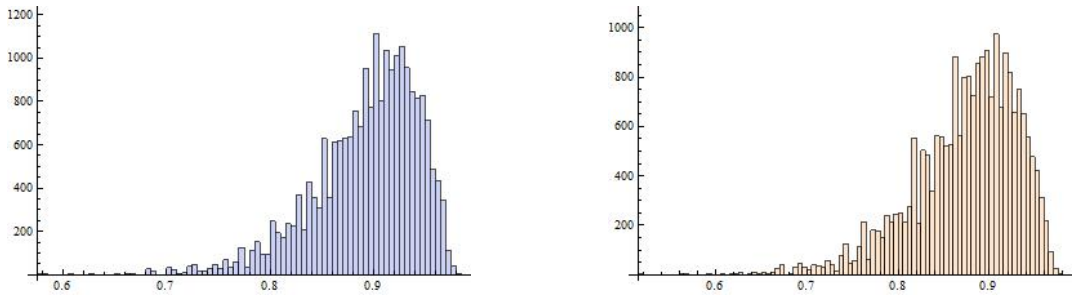
Tabela 4.20: PPagas: Ajustamento da *Proporção das Prestações Pagas*

Mostra-se seguidamente alguns exemplos da *Proporção das Prestações Pagas* para clientes incumpridores com diferentes características, clientes definidos na Tabela 4.10,tendo em conta os critérios inicialmente escolhidos, desfazendo a mudança de variável.

Cliente	1° Critério	2° Critério
a	0,945046	0,943965
b	0,949972	0,930410
c	0,959008	0,878323
d	0,925292	0,931641
e	0,896458	0,899474

Estimou-se, também, a *Proporção das Prestações Pagas* da carteira de crédito ao consumo da Instituição Bancária de Cabo Verde, segundo os dois critérios definidos. Na Figura 4.20 pode-se observar a distribuição da *Proporção das Prestações Pagas* da carteira, para ambos os critérios.

	b_i
1° Critério	0,89298
2° Critério	0,87447



(a) 1° Critério (b) 2° Critério

Figura 4.20: Distribuição da *Proporção das Prestações Pagas* da carteria

4.4 Taxa de Recuperação do Crédito - Regressão Beta

A *Taxa de Recuperação* é, por vezes, um tema negligenciado na análise de risco, mas tal como visto no capítulo 2, é uma medida de importância relevante. Tendo em conta que é uma medida que permite mensurar o montante de crédito não recuperado, para clientes que não liquidam por completo os seus empréstimos. Assim, é fundamental que uma instituição bancária estime a taxa de recuperação da carteira ou por cliente, ou seja, que obtenha uma estimativa da proporção do montante a recuperar, e caso de ocorrência de *default*.

É comum estudar-se um outro conceito, a taxa de perda em caso de incumprimento por parte de um cliente, sendo esta taxa definida como *Loss Given Default* - LGD - que na realidade é o complementar da taxa de recuperação ($L = 1 - R$).

4.4.1 Ajustamento de dados - Taxa de Recuperação

Neste estudo, para se estimar a taxa de recuperação, utilizou-se uma variável *proxy* da variável de interesse, uma vez que não foram disponiveis dados suficientes para calcular a taxa de recuperação para cada cliente da base de dados. Assim, calculou-se a proporção em dívida relativamente ao montante em dívida no vencimento do contrato, ou seja, o quociente entre o montante em dívida e o valor de empréstimo do contrato. Resumidamente, para se estimar a taxa de recuperação, ir-se-á estimar primeiramente a taxa de perda acima definida, a LGD.

O método para estimação da taxa de recuperação é análogo ao usado na secção anterior, a Regressão Beta. Assim, é considerado o seguinte conjunto de variáveis:

Idade	Agência
Género	V. Empréstimo
Estado Civil	Prazo
Habilitações	Tx. Nominal
Act. Profissional	V. Prestação
Ent. Patronal	T. Garantia

Ajustando uma Regressão Beta ao modelo completo e sendo a *Loss Given Default* a variável resposta, obteve-se os seguintes resultados:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.789594	0.120661	-6.544	5.99e-11	***
TxN2	0.005409	0.049071	0.110	0.912227	
VPrest2	-0.069403	0.053761	-1.291	0.196714	
VPrest3	-0.106633	0.069037	-1.545	0.122446	
VPrest4	-0.075310	0.084735	-0.889	0.374124	
VEmprest2	-0.017146	0.050256	-0.341	0.732966	
VEmprest3	0.188390	0.070122	2.687	0.007218	**
VEmprest4	0.201505	0.103068	1.955	0.050575	.
Idade2	0.085352	0.062584	1.364	0.172626	
Idade3	0.039236	0.052603	0.746	0.455740	
Idade4	0.036657	0.047073	0.779	0.436145	
Idade5	-0.016108	0.050526	-0.319	0.749878	
Prazo2	-0.204185	0.057064	-3.578	0.000346	***
Prazo3	-0.281901	0.058535	-4.816	1.46e-06	***
Prazo4	-0.242147	0.065122	-3.718	0.000201	***
Prazo5	-0.282060	0.074331	-3.795	0.000148	***
Agencia2	-0.311043	0.052275	-5.950	2.68e-09	***
Agencia3	-0.459549	0.074444	-6.173	6.70e-10	***
Agencia4	-0.355903	0.044452	-8.006	1.18e-15	***
Agencia5	-0.297555	0.050916	-5.844	5.09e-09	***
ActProfissional2	-0.191000	0.049241	-3.879	0.000105	***
ActProfissional3	-0.291640	0.076453	-3.815	0.000136	***
ActProfissional4	-0.178332	0.053947	-3.306	0.000947	***
GeneroM	-0.012486	0.030881	-0.404	0.685970	
EntPatronal2	-0.042021	0.069101	-0.608	0.543117	
EntPatronal3	0.036694	0.077844	0.471	0.637369	
EntPatronal4	0.191072	0.074984	2.548	0.010829	*
EntPatronal5	0.172887	0.074969	2.306	0.021104	*
Civil2	0.068442	0.039374	1.738	0.082169	.
Habilitacoes2	-0.021641	0.040841	-0.530	0.596199	
Habilitacoes3	-0.050829	0.052481	-0.969	0.332788	
Habilitacoes4	0.002164	0.080596	0.027	0.978575	
Habilitacoes5	0.113725	0.081235	1.400	0.161530	
Garantia2	0.172940	0.033236	5.203	1.96e-07	***

Tabela 4.21: Proporção em Dívida - Modelo Completo

De seguida, realizou-se a mesma análise para o modelo nulo e posteriormente, comparou-se os dois modelos com o propósito de escolher o modelo que melhor estima a variável resposta. Para se comparar os modelos, com o fim de seleccionar o melhor modelo, observou-se os valores do Critério de Informação de Akaike (AIC), sendo o valor de AIC para o modelo completo de $-4.863,032$ e para o modelo nulo de $-4.671,302$. Logo deve-se rejeitar a hipótese de que o modelo nulo é estatisticamente melhor que o modelo completo, pelo que é aceitável afirmar que os coeficientes acrescidos são significativamente diferentes de zero.

Seguindo a analogia apresentada nos casos anteriores, tanto para estimação da probabilidade de *default* como para a estimação da *Proporção de Prestações Pagas*, utilizando o método *Stepwise* para seleccionar as covariáveis a retirar do modelo e o valor AIC para comprar os modelos, obteve-se o melhor modelo de ajustamento para a variável em estudo, a proporção em dívida (Tabela 4.22).

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.89077	0.07722	-11.536	< 2e-16	***
VEmprest3	0.14597	0.03449	4.233	2.31e-05	***
VEmprest4	0.15928	0.04809	3.312	0.000927	***
Prazo2	-0.19697	0.05517	-3.570	0.000356	***
Prazo3	-0.24691	0.04987	-4.951	7.38e-07	***
Prazo4	-0.21407	0.05401	-3.964	7.38e-05	***
Prazo5	-0.24742	0.06132	-4.035	5.46e-05	***
Agencia2	-0.29837	0.05152	-5.792	6.97e-09	***
Agencia3	-0.45496	0.07359	-6.182	6.31e-10	***
Agencia4	-0.35017	0.04402	-7.954	1.80e-15	***
Agencia5	-0.28768	0.05062	-5.683	1.32e-08	***
ActProfissional2	-0.20154	0.04643	-4.341	1.42e-05	***
ActProfissional3	-0.30892	0.07355	-4.200	2.67e-05	***
ActProfissional4	-0.19344	0.04897	-3.950	7.81e-05	***
EntPatronal4	0.21011	0.04131	5.086	3.65e-07	***
EntPatronal5	0.19501	0.04052	4.812	1.49e-06	***
Civil2	0.08559	0.03602	2.376	0.017489	*
Habilitacoes5	0.14672	0.07331	2.001	0.045354	*
Garantia2	0.17723	0.03307	5.360	8.32e-08	***

Tabela 4.22: Proporção em Dívida - Modelo de ajustamento final

4.4.2 Análise de Resíduos

Finalizado o ajustamento dos dados em relação à taxa de recuperação, deve-se testar a adequabilidade dos modelos encontrados e esta é verificada através da análise dos resíduos.

A análise de resíduos foi efectuada com base nos resíduos denominados por Desvio Residual, concluindo-se que os pontos apresentam um desvio residual padronizado de valor compreendido entre $[-2, 4]$, como se pode observar na Figura 4.21, bem como um gráfico com os quantis.

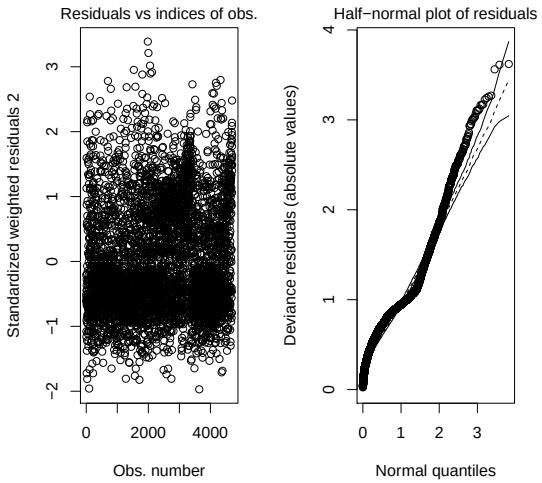


Figura 4.21: Taxa de Recuperação - Análise de Resíduos

Quanto à análise de pontos influentes, recorreu-se à distância de Cook e considerou-se observações influentes aquelas cuja distância de Cook é superior a $0,005$, ou seja, as observações que possuem as distâncias de Cook mais elevadas. Pelo que, foram excluídas as observaçõs definidas como influentes.

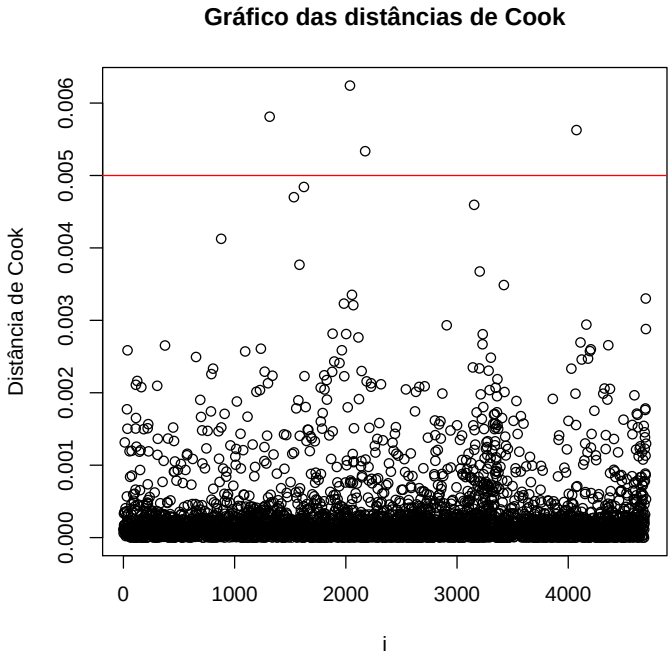


Figura 4.22: Taxa de Recuperação - Distâncias de Cook

4.4.3 Estimação da Taxa de Recuperação

Concluído o estudo estatístico do modelo de ajustamento da taxa de recuperação, a análise de resíduos respectiva e a exclusão das observações discordantes, apresenta-se os resultados finais da estimação da proporção em dívida de um dado cliente definido como incumpridor, LGD_i, definida no capítulo anterior e que pode ser escrita da seguinte forma:

$$LGD_i = \frac{exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1+exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}$$

onde, $Y_i = \beta_0 + \beta_1 X_1 + ... \beta_p X_p$ com $Y \sim \mathcal{B}(\mu, \phi)$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^T$ é um vector de parâmetros de regressão desconhecidos e $\boldsymbol{x_i} = (x_{i1}, ..., x_{ik})^T$ representa o vector de covariáveis.

Para esta regressão, entende-se como cliente padrão aquele que assume as seguintes características:

Variável	Grupo
Civil1	Casado, Divorciado, Separado
Habilitacoes1,2,3e4	Sem curso Superior
ActProfissional1	Doméstica, Estudante, Peq./Med. Empresa
EntPatronal1,2e3	Inst. Financeiras, Aposentado/Pensionista, C. Municipal e outros
Agencia1	4, 7, 14, 18, 28 e 29
Prazo1	Inferior a 13
VEmprest2e3	< 200.000
Garantia1	Depósitos, Hipotecas e Outros

Tabela 4.23: Modelo de regressão para LGD: Cliente Padrão

Tendo como cliente padrão, o cliente que assume as características acima descritas, estimou-se a que a taxa de recuperação do cliente padrão é de:

LGD	R
0,290951	0,709049

Seguidamente, na Tabela 4.24, apresenta-se os valores estimados para a proporção em dívida de um cliente incumpridor, após remoção das observações discordantes.

Características	Coeficientes
(Intercept)	-0.89077
VEmprest3	0.14597
VEmprest4	0.15928
Prazo2	-0.19697
Prazo3	-0.24691
Prazo4	-0.21407
Prazo5	-0.24742
Agencia2	-0.29837
Agencia3	-0.45496
Agencia4	-0.35017
Agencia5	-0.28768
ActProfissional2	-0.20154
ActProfissional3	-0.30892
ActProfissional4	-0.19344
EntPatronal4	0.21011
EntPatronal5	0.19501
Habilitacoes5	0.14672
Civil2	0.08559
Garantia2	0.17723

Tabela 4.24: Modelo de regressão para LGD

Mostra-se seguidamente alguns exemplos da taxa de recuperação para clientes incumpridores com diferentes características, clientes definidos na Tabela 4.10.

Cliente	LGD	R
a	0,2772	0,7228
b	0,2179	0,7821
c	0,2748	0,7252
d	0,2454	0,7546
e	0,2975	0,7025

Estimou-se, ainda a taxa de recuperação da carteira de crédito ao consumo, sendo esta de, $78,1459\%$. Na Figura 4.23 pode-se observar a distrituição da *Taxa de Recuperação* da carteira.

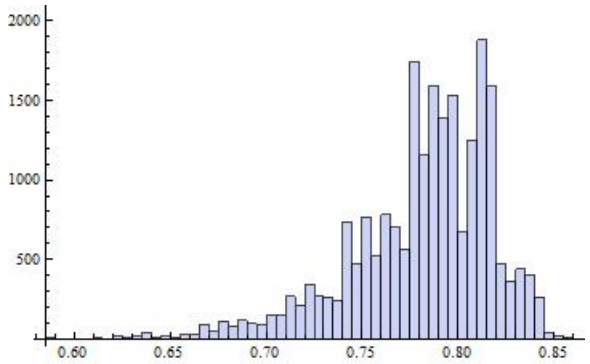


Figura 4.23: Distrituição da *Taxa de Recuperação* da carteira

4.5 Estimação do *Spread* - Metodologia Actuarial

Nesta secção será calculado o *spread* da carteira segundo o modelo apresentado na secção 1.2.3 e proposto em [EGFS14]. Mas antes de se expor o resultado, ir-se-á apresentar os resultados obtidos para a probabilidade de *default* e para a taxa de recuperação segundo a notação apresentada no modelo, para que a leitura seja mais coerente. Os resultados serão apresentados apenas para o primeiro critério definido, uma vez que é o mais usual na literatura.

Considere-se que Λ e Δ são variáveis aleatórias, que representam a taxa de recuperação e a probabilidade de *default* estimadas, para cada cliente da carteira, nas secções anteriores, e que $\Lambda \cdot X_T = F(X_\tau, \cdot)$, conforme definido na secção 1.2.3. Note-se que para a estimação da taxa de recuperação foi usada uma *proxy* desta variável, a proporção em dívida no vencimento da maturidade do contrato, por falta de disponibilidade de dados e informação para a definição da taxa de recuperação. Para a probabilidade de *default*, também foi utilizada umaa variável *proxy*, definida como o primeiro critério de incumprimento (ver Tabela 4.3).

Assumindo, que Λ e X_T (montante de empréstimo concedido) são independentes, que a probabilidade de *default* da carteira é dada por $\mathbb{E}[\Delta] = \mathbb{P}[\tau \leq T]$ e que a taxa de recuperação é dada por $\lambda = \mathbb{E}[\Lambda]$, então a fórmula em (2.3) é válida, mostrando-se que os *proxies* adoptados recaem sobre os pressupostos do modelo.

Das análises efectuadas obtiveram-se as seguintes estimativas:

- Probabilidade de *default*: $\mathbb{E}[\Delta] = \mathbb{P}[\tau \leq T] = 0,108287$
- Taxa de Recuperação: $\lambda = \mathbb{E}[\Lambda] = 0,781459$
- *Spread*: $s_T = 0,0242387$

Como referido em [EGFS14], mostra-se que, para os dados constantes na carteira de crédito utilizada, se verifica a independência entre as variáveis Λ e Δ, pelo que se demonstra seguidamente o teste de independência realizado, como objectivo de validar o pressuposto.

Para testar a hipótese de que se as variáveis são ou não independentes, utilizou-se o teste de independência do Qui-Quadrado (χ^2_k de Pearson), teste este que permite averiguar se as variáveis estão relacionadas. A hipótese nula afirma que as variáveis são independentes e a sua estatística de teste é definida como:

$$\chi^2_k = \sum_{i=1}^{n} \frac{(O_j - E_j)^2}{E_j} ,$$

com k graus de liberdade e E_j e O_j as frequências esperadas e as frequência observadas para cada classe j, $j = 1, ..., n$ e n dimensão da carteira.

Pelo que se obteve $\chi^2_k = 1874,94$, com $k = 22044 - 1 = 22043$ graus de liberdade. E para este teste obteve-se um *p-value* muito próximo de 1, donde, ao nível de significância de 5%, por exemplo, se rejeita a hipótese de que as variáveis não são independentes. Assim, é aceitável considerar-se que as variáveis Λ e Δ são independentes.

Assim, aplicando as equações (2.6) e (2.7), obteve-se como estimativa de $\mathbb{E}[s_{\text{cliente}}]$ a média da carteira $\overline{s_{\text{cliente}}} = 0,0247791$ e $\mathbb{E}[1 - \Lambda]\,\mathbb{E}[\Delta] = 0,0236651$, pelo que é razoável definir um spread individualmente para cada cliente, utilizando a taxa de recuperação, Λ, e probabilidade de *default*, Δ.

Mostra-se, em seguida, alguns exemplos de *spread* para clientes, com as características que foram usadas em exemplos anteriores (ver Tabela 4.10).

Cliente	Δ	Λ	s
a	0,2954	0,7228	0,089188
b	0,3376	0,7821	0,079404
c	0,0898	0,7252	0,025301
d	0,1053	0,7546	0,026526
e	0,2750	0,7025	0,089102

Tabela 4.25: Exemplos - Estimação do *spread*

Na Tabela 4.25, para diferentes clientes observam-se valores de *spread* distintos, como por exemplo, ao cliente a foi atribuído, aproximadamente, um *spread* de $8,9\%$, enquanto que para o cliente c foi atribuído um spread de $2,53\%$, pelo que se pode concluir que conhecer as características de cada cliente e do contrato de empréstimo é proponente na definição do risco/*spread*.

Nesta secção foi proposto um modelo simples que permite a estimação do *spread* da carteira de crédito ao consumo, em função da probabilidade de *default* e da taxa de recuperação dos contratos de crédito. E, ainda, se mostrou que é possível definir os *spreads* individuais de cada cliente de uma forma coerente, pelo que o modelo pode ser utilizado para determinar uma medida do risco de crédito para novos clientes da carteira.

5

Conclusão

Esta dissertação teve como primeiro objectivo estimar a probabilidade de incumprimento, da carteira e do cliente de uma carteira de crédito ao consumo de uma Instituição Bancária de Cabo Verde. Assim, estimou-se a probabilidade de incumprimento de acordo com dois critérios para cliente incumpridor, sendo o primeiro através do número de dias de incumprimento durante o prazo do contrato e o segundo critério através da existência de valor em dívida no vencimento do processo de empréstimo. Para a estimação da probabilidade de *default* utilizou-se a Regressão Logística para a estimação das probabilidades e as técnicas *Stepwise-Backward* e *AIC* para a selecção das variáveis mais significativas para explicar a ocorrência de *default*. Pelo que se obteve para a probabilidade de *default* da carteira de 0,108287, segundo o primeiro critério e de 0,193863, utilizando a definição do segundo critério, sendo esta última superior. Na Tabela 4.13, da página 56, encontram-se exemplos da estimação da probabilidade de *default* para alguns clientes e verifica-se que para alguns clientes a probabilidade de default é superior no primeiro critério adoptado.

Estudou-se, também, a proporção de prestações pagas por um cliente classificado como incumpridor, considerando os dois critérios acima descritos e ajustando uma Regressão Beta, o que permitiu facultar uma análise complementar. Tendo sido possível estimar o número de prestações pagas durante um contrato de crédito que entre em *default*, uma vez que o número de prestações totais é estipulado no início do contrato. Pelo que se estimou, para o primeiro critério, uma proporção das prestações pagas da carteira de 0,89298 e considerando o segundo critério de 0,87497, ambos valores elevados.

O segundo objectivo da dissertação teve como base estimar a taxa de recuperação tanto da carteira como do cliente, utilizando a Regressão Beta, de forma a obter, em caso

de ocorrer *default*, uma percentagem do montante de crédito concedido que a instituição bancária poderá vir a recuperar. Esta medida é importante uma vez que poderá evitar insolvência das instituições bancárias e permite quantificar mais adequadamente o risco envolvido num contrato de crédito. Concluiu-se que a carteira de crédito ao consumo, possui uma taxa de recuperação de $78,1459\%$.

Por fim, propondo um modelo a tempo discreto, publicado ao longo do desenvolvimento desta dissertação, ver [EGFS14], estimou-se o *spread* da carteira em função da probabilidade de incumprimento, estimada através das variáveis sócio-demográficas e descritivas dos contratos de crédito e da taxa de recuperação de cada cliente e ainda foi possível definir individualmente o spread de cada cliente.

O presente trabalho permitiu a consolidação dos conhecimentos académicos adquiridos sobre os Modelos Lineares Generalizados e deu a conhecer os modelos de análise de risco como sendo de máxima importância para as instituições bancárias. Conclui-se que se o risco de incumprimento de um cliente for devidamente analisado e previamente, bem como o *spread* adequado que deve ser cobrado para fazer face ao risco, esta análise pode propocionar lucro à instituição bancária.

Fica em aberto algumas questões, que não foram analisadas neste trabalho, mas que podem ser objecto de estudos futuros, nomeadamente: a estimação da probabilidade de *default* segundo outros critérios, diferentes dos estipulados; em vez de se estimar a taxa de recuperação através de uma *proxy* da variável, utilizar o seu valor real; estimar o spread através de outros modelos existentes na literatura, como por exemplo, modelos financeiros.

Bibliografia

[AMR11] S. Ahmad, H. Midi e N. Ramli. *Diagnostics for Residual Outliers Using Deviance Component in Binary Logistic Regression*. World Applied Sciences Jownal 14 (8): 1125-1130, 2011, 2011.

[And04] F. Andrade. *Desenvolvimento de Modelo de Risco de Portfólio para Carteiras de Crédito a Pessoas Físicas*. Tese de Mestrado, Escola de Admnistração de Empresas de São Paulo - Fundação Getúlio Vargas, 2004.

[ACn] E. Araujo e C. Carmona. *Construção de Modelos de Credit Scoring com análise discriminante e Regressão Logística para a gestão do Risco de inadimpência de uma instituição de Microcrédito*. Read - edição 62 vol 15 n°1, Jan-Abr 2009.

[BH10] E. Brian e T. Hothorn. *Handbook of Statistical Analysis using R*. Chapman & Hall / CRC, 2^a edição, USA, 2010.

[BN09] G. Brito e A. Neto. *Modelo de Classificação de Risco de Crédito de Empresas*. Revista Contemporânea de Contabilidade, Volume 19, n°46, pág. 18-29, 2009.

[Cae11] V. Caeiro. *Avaliação do Risco de Crédito de Clientes Empresariais, Levantamento de Requisitos e Estimação de Modelos*. Tese de Mestrado, Instituto Superior de Estatística e Gestão - Universidade Técnica Lisboa, 2011.

[CAN98] J. Caouette, E. Altman e P. Narayanan. *Managing Credit Risk: The Next Great Financial Challenge*. Hardcover, 1998.

[Cha03] A. Chaia. *Modelos de Gestão de Risco do Crédito e a sua Aplicabilidade no Mercado Brasileiro*. Tese de Mestrado - Universidade de São Paulo, 2003.

[Cra02] J. Cramer. *The Origins of Logistic Regression*. Tinbergen Istitute Discussion Paper, 2002.

[Cra07] M. Crawley. *The R book*. John Wiley & Sons, Inc., 2007.

[CNZ10] F. Cribari-Neto e A. Zeileis. *Beta Regression in R*. Journal of Statistical Software, Volume 34, Issue 2, 2010.

[Cru12] J. Cruz. *Cálculo da Loss Given Default no Crédito à Habitação com Cadeias de Markov*. Tese de Mestrado - Instituto Superior de Estatística e Gestão - Universidade Técnica Lisboa, 2012.

[Dob02] A. Dobson. *An Introdution to Generalized Linear Models*. Chapman & Hall / CRC, 2002.

[EGFS14] M. L. Esquível, G. R. Guerreiro, J. M. Fernandes e A. F. Silva. *On a Spread Model for Portfolio Credit Risk Modeling*. Proceedings of Conference ICNAAM, Greece, 2014.

[Fer12] J. M. Fernandes. *Estudo de uma Carteira de Crédito ao Consumo de um Banco de Cabo Verde*. Tese de Doutoramento, Instituto Superior de Estatística e Gestão de Informação - Universidade Nova de Lisboa, 2012.

[FCN04] S. Ferrari e F. Criabari-Neto. *Beta Regression for Modelling Rates and Proportions*. Journal of Applied Statistics, Volume 31, Issue 7, 2004.

[Fig06] C. Figueira. *Modelos de Regressão Logística*. Tese de Mestrado, Universidade Federal do Rio Grande do Sul - Instituto de Matemática, 2006.

[Fin03] Finder. *Rating de Risco de Crédito*. Disclosure das Transações de Crédito, edição n°92, Ano VIII, 2003.

[Gey12] C. J. Geyer. *The Wilks, Wald, and Rao Tests*. Stat 8112 Lecture Notes, 2012.

[GGM13] E. Gonçalves, M. Gouvês e D. Mantovani. *Análise de Risco de Crédito com o uso de Regressão Logística*. Revista Contemporânea de Contabilidade, Volume 10, n°20, pág. 139-160, 2013.

[GJZ09] X. Guo, J. Jarrow e Y. Zeng. *Modelling the Recovery Rate in a Reduced form Modell*. Mathematical Finance, Volume 19, n°1, pág. 73-97, 2009.

[HBRA10] J. Hair, W. Black, B. Robin e R. Anderson. *Multivariate Data Analysis*. Prentice Hall, 2010.

[HLS13] D. Hosmer, S. Lemeshow e R. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, Inc., New Jersey, 2013.

[JLT97] J. Jarrow, D. Lando e S. Turnbull. *A Markov Model for the Term Structure of Credit Risk Spreads*. The Review of Financial Studies Summer, Volume 10, n°2, pág. 481-523, 1997.

[Lew92] E. Lewis. *An Introduction to Credit Scoring*. Athena Press, California, 1992.

[MN89] P. McCullagh e J. Nelder. *Generalized Linear Models*. Chapman & Hall / CRC, USA, 1989.

[MFE05] A. J. McNeil, R. Frey e P. Embrechts. *Quantitative Risk Management*. Princeton University Press, New Jersey, 2005.

[Mes97] L. Mester. *What's the point of credit scoring*. Business Review, Federal Reserve Bank of Philadelphia, 1997.

[MSP02] B. Murteira, C. Silva e J. Pimenta. *Introdução à Estatística*. McGraw-Hill, Portugal, 2002.

[NW72] J. Nelder e R. Wedderburn. *Generalized Linear Models*. Journal of the Royal Statistical Society, 1972.

[Nun11] A. Nunes. *Modelação Espacial de Acidentes Rodoviários na Cidade de Lisboa*. Tese de Mestrado, Faculdade de Ciências e Tecnologias - Universidade Nova de Lisboa, 2011.

[RS01] V. Rohatgi e A. Saleh. *An Introduction to Probability and Statistics*. John Wiley & Son, Inc., 2001.

[SA02] A. Saunders e L. Allen. *Credit Risk Measurement: New Approaches To Value-At-Risk and other Paradigms*. Handcover, 2002.

[Sec02] J. Securato. *Crédito: Análise e Avaliação do Risco: Pessoas Físicas e Jurídicas*. São Paulo, 2002.

[Sem09] D. Semedo. *Credit Scoring: Aplicação da Regressão Logística vs Redes Neuronais Artificiais na Avaliação do Risco de Crédito no Mercado Cabo-Verdiano*. Tese de Mestrado, Instituto Superior de Estatística e Gestão de Informação - Universidade Nova de Lisboa, 2009.

[Sid06] N. Siddiqui. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. John Wiley & Sons, Inc., New Jersey, 2006.

[TS00] M. Turkman e J. Silva. *Modelos Lineares e Generalizados da teoria à prática*. Edições SPE, Lisboa, 2000.

[Val10] C. Vale. *Modelação e Estimação do Risco de Crédito - Estudo de uma Carteira*. Tese de Mestrado, Faculdade de Ciências e Tecnologias - Universidade Nova de Lisboa, 2010.